JSCE (Journal of System and Computer Engineering)
Vol.7, No.4, October 2026, pp. 408-419
ISSN (online): 2723-1240
DOI : https://doi.org/10.61628/jsce.v7i4.2850

Research Article Open Access (CC–BY-SA) ■

Fine-Tuning Whisper Small for Low-Resource Makassar Automatic Speech Recognition

Hendra Sagita Putra [1,a,*]; Ismail Gaffar [2,b]; Amiruddin A. [3,c]; Muh Risfandi Samad [4,d]; Akbar Hasrun [5,e]; Anas [6,f]; Tirta Chiantalia Sharief [7,g];

[1] *Universitas Tamalatea Makassar, Makassar, Indonesia*
[2] *Politeknik Pertanian Negeri Pangkajene Kepulauan, Pangkajene & Kepulauan, Indonesia*
[3] *Politeknik Lembaga Pendidikan dan Pengembangan Profesi Indonesia Makassar, Makassar Indonesia*
[4,6,7] *Universitas Muhammadiyah Barru, Barru, Indonesia*
[5] *Universitas Handayani Makassar, Makassar, Indonesia*
[a] *Hendra.sagita@utama.ac.id*; [b] *ismail.gaffar@polipangkep.ac.id*; [c] *amiruddinardinmks@gmail.com*;
[d] *risfandi.samad8@gmail.com*; [e] *akbarhasrun94@gmail.com*; [f] *anas.daeng.pasewang@gmail.com*;
[g] *tirtashariff@gmail.com* ;
** Corresponding author*

Abstract

Automatic speech recognition for regional languages is constrained by limited paired speech-text data and evaluation settings that may obscure differences between familiar and unseen transcript content. This study evaluates full-model fine-tuning of Whisper Small for Makassar using 6,119 utterances totaling 6.3882 hours. The corpus comprised 4,894 training, 612 validation, and 613 locked test utterances; the same three speakers occurred in all partitions, making the evaluation speaker-dependent. Zero-shot and fine-tuned models were re-evaluated on identical test items using matched preprocessing, 16-kHz audio, Indonesian transcription settings, deterministic one-beam decoding, and a 225-token generation cap. Overall WER decreased from 120.84% to 51.78%, while CER decreased from 62.00% to 16.19%. A post-hoc content audit identified 136 seen-content and 477 unseen-content test utterances. On unseen content, WER decreased from 125.88% to 60.91%. Error decomposition showed substitutions remained dominant after fine-tuning, while targeted zero-shot cases exhibited repetitive long-generation behavior. The findings demonstrate substantial corpus-specific adaptation gains while emphasizing that aggregate error rates should be interpreted alongside transcript-content overlap, speaker design, decoding behavior, and reproducibility constraints.

Keywords—*automatic speech recognition, Makassar language, Whisper, low-resource speech recognition, reproducibility*

1. Introduction

Automatic speech recognition (ASR) has advanced substantially through large pretrained speech models, but these advances remain unevenly distributed across languages. Low-resource languages are commonly constrained by limited labeled speech and other language resources, which complicate both model development and reliable evaluation (Alharbi et al., 2021; Slam et al., 2023). In such settings, performance should not be interpreted only from a single aggregate error rate because corpus composition, transcript familiarity, speaker coverage, and inference conditions can materially affect the resulting measurement. Large multilingual resources such as Common Voice and FLEURS illustrate the growing scale of multilingual speech data and provide explicit benchmarks for evaluating speech technology across many languages (Ardila et al., 2020; Conneau et al., 2023).

Whisper is a multilingual encoder–decoder Transformer pretrained on approximately 680,000 hours of weakly supervised multilingual and multitask audio (Radford et al., 2023). Its large-scale

pretraining provides a useful starting point for low-resource ASR, but pretrained multilingual coverage does not guarantee strong zero-shot recognition for every regional language. Recent studies have therefore investigated target-language adaptation of Whisper in low-resource settings. Liu et al. (2024), for example, examined multiple Whisper fine-tuning strategies across low-resource languages, while parameter-efficient approaches have also been investigated using LoRA and related adaptation mechanisms (Polat et al., 2024; Ghimire et al., 2024). Studies involving Indonesian and regional-language settings further demonstrate the growing interest in adapting pretrained ASR models to linguistically underrepresented speech (Pratama & Amrullah, 2024; Jafar et al., 2025). Collectively, these studies establish that target-language adaptation can substantially improve recognition, although the resulting performance depends on the language, corpus, model configuration, adaptation strategy, and evaluation protocol. The broader parameter-efficient adaptation literature provides established foundations for adapter modules and low-rank adaptation (Houlsby et al., 2019; Hu et al., 2022)while recent Whisper studies continue to explore LoRA-based optimization under low-resource conditions (Wang et al., 2025).

Beyond Whisper-specific adaptation, self-supervised multilingual speech research has established strong foundations for low-resource ASR. wav2vec 2.0 demonstrated that self-supervised pretraining can substantially reduce dependence on labeled speech (Baevski et al., 2020)while XLS-R extended cross-lingual representation learning to large multilingual settings (Babu et al., 2022). The Massively Multilingual Speech (MMS) project further scaled speech technology to more than one thousand languages and used speech derived from readings of publicly available religious texts as part of its multilingual resource construction (Pratap et al., 2024). In the Indonesian context, Sakti and Titalim evaluated self-supervised models using multilingual Indonesian ethnic-language data, reinforcing the relevance of pretrained multilingual speech models for regional low-resource ASR (Sakti & Titalim, 2023).

Recent work has also begun to examine Makassar speech directly. Azis et al. evaluated Whisper Tiny, Base, and Small together with a Wav2Vec2/XLS-R model on a small spontaneous-speech Makassar benchmark (Azis et al., 2026). That study provides important evidence that Makassar ASR is emerging as a distinct evaluation setting. The present study addresses a different experimental question: it evaluates full-model adaptation of Whisper Small on a larger processed read-speech corpus and explicitly separates seen from unseen transcript content under a controlled, speaker-dependent protocol. Accordingly, the contribution claimed here is complementary and evaluation-centered rather than a claim of being the first Makassar ASR study.

An important evaluation question nevertheless remains when adaptation is performed on a very small regional-language corpus. A reduction in aggregate word error rate (WER) does not by itself indicate whether the improvement extends to transcript content that was not represented during training. The interpretation becomes even more restricted when the same speakers occur across training and test partitions, because performance under such a design cannot be treated as evidence of unseen-speaker or population-level generalization. Consequently, adaptation gains in small corpora need to be interpreted together with transcript-content overlap, speaker composition, error characteristics, and the reproducibility of the inference protocol rather than through aggregate WER alone.

This issue is particularly relevant to the Makassar corpus examined in this study. The available processed corpus contains only three speakers, and all three occur across the training, validation, and test partitions. In addition, a post-hoc transcript-content audit shows that part of the locked test set shares normalized transcript content with the training data. The central research problem is therefore not merely whether full-model fine-tuning reduces recognition error, but whether the observed improvement remains evident on unseen transcript content and how that improvement should be interpreted within a speaker-dependent evaluation design.

Accordingly, this study aims to evaluate the effect of full-model fine-tuning of Whisper Small for low-resource Makassar ASR under a controlled evaluation protocol and to examine how the resulting performance differs between seen and unseen transcript content while accounting for

speaker composition, word-error characteristics, and reproducibility constraints. The novelty of the study lies not in proposing a new fine-tuning algorithm, but in providing an evaluation-centered analysis of Whisper Small adaptation for a very low-resource Makassar corpus. Rather than relying solely on aggregate pre- and post-fine-tuning error rates, the study combines controlled model comparison with transcript-content visibility analysis, explicit speaker-dependent interpretation, word-error decomposition, and a reproducibility audit.

The contributions of this study are fourfold. First, it provides a controlled comparison between zero-shot and full-model fine-tuned Whisper Small on the identical locked 613-utterance test set using matched preprocessing, normalization, language/task configuration, and deterministic decoding conditions. Second, it performs a transcript-content audit that separates 136 seen-content from 477 unseen-content test utterances, enabling the adaptation effect to be examined beyond aggregate WER and beyond transcript content represented during training. Third, it combines word-level substitution, deletion, and insertion analysis with an explicit characterization of the experiment as speaker-dependent, thereby preventing unseen-content performance from being conflated with unseen-speaker generalization. Fourth, it conducts a reproducibility-oriented audit that separates historical measurements from newly controlled results and examines the influence of decoding behavior on the interpretation of zero-shot performance. Together, these contributions provide a more transparent account of what full-model Whisper Small adaptation achieves—and does not establish—under the available low-resource Makassar corpus.

2. Method

2.1 Research Design

The study used an experimental fine-tuning and controlled re-evaluation design. Makassar audio–transcript pairs were prepared for Whisper, Whisper Small was fine-tuned as a full model, and the selected checkpoint was evaluated on a locked final test set. After the original experiment, reviewer-driven audits examined transcript-content overlap, error composition, historical reproducibility, and zero-shot generation behavior. No additional training was performed during the final controlled baseline audit. Figure 1 summarizes the complete methodological workflow and the boundary between the original experimental procedure and the subsequent reviewer-driven audits.

2.2 Dataset Source and Corpus Construction

The experimental corpus was constructed from selected Makassar-language New Testament recordings obtained from the publicly accessible SABDA Audio platform (https://audio.sabda.org/). SABDA's copyright information identifies the Makassar Bible text as © 1999 Lembaga Alkitab Indonesia (LAI) and the corresponding audio recording as ℗ 2012 Hosanna. The study did not use the complete source collection. Only selected recordings that passed the project's preparation, segmentation, transcript-alignment, and filtering stages were included in the experimental corpus. The source recordings are third-party copyrighted materials and are not redistributed as part of this study; access to the original recordings should be obtained from the source platform and remains subject to the applicable rights and usage conditions. The use of read religious-text speech as a multilingual ASR resource also has precedent in MMS, although that precedent does not determine the licensing conditions of the present third-party source material (Pratap et al., 2024).

After processing, the effective corpus contained 6,119 utterances with a total duration of 6.3882 hours. The official partition comprised 4,894 training utterances (79.98%), 612 validation utterances (10.00%), and 613 test utterances (10.02%), corresponding to approximately 5.1597, 0.6060, and 0.6224 hours, respectively. Across the complete corpus, the mean utterance duration was 3.7583 s and the median was 3.0 s; durations ranged from 0.12 to 30.0 s. The corpus contained 41,667 transcript words, averaging 6.8094 words per utterance.

The same three speaker identifiers (Lukas, Markus, and Matius) occur in the training, validation, and test partitions. Consequently, the final evaluation is speaker-dependent. The study does not claim generalization to unseen speakers.

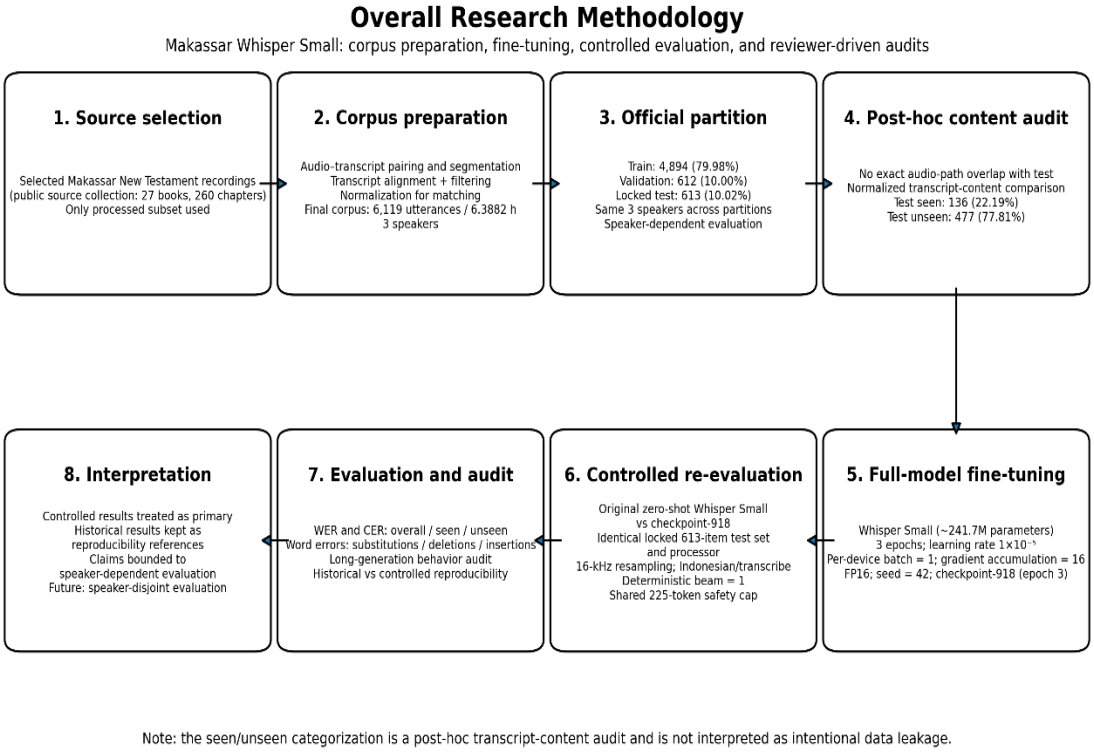


Figure 1. Overall research methodology for the Makassar Whisper Small study, from corpus preparation and official partitioning through full-model fine-tuning, controlled zero-shot versus fine-tuned re-evaluation, post-hoc transcript-content analysis, and reproducibility auditing.

Table 1. Characteristics of the processed Makassar speech corpus.

Characteristic	Value
Total processed corpus	6,119 utterances
Effective audio duration	6.3882 h
Training partition	4,894 utterances (79.98%); 5.1597 h
Validation partition	612 utterances (10.00%); 0.6060 h
Final test partition	613 utterances (10.02%); 0.6224 h
Speakers	3 (Lukas, Markus, Matius)
Evaluation design	Speaker-dependent
Mean / median duration	3.7583 s / 3.0 s
Duration range	0.12–30.0 s
Transcript words	41,667
Audio format	WAV, mono, PCM-16/16-bit; source sampling rate 11,025 Hz

2.3 Dataset Split and Post-hoc Content Audit

No exact audio-path overlap was found between training and test or between validation and test. Separately, a post-hoc transcript-content audit was conducted using the historically reconstructed normalization procedure. A test item was categorized as seen-content when its normalized reference transcript was also present in training; otherwise it was categorized as unseen-content. Of the 613 test utterances, 136 (22.19%) were seen-content and 477 (77.81%) were unseen-content. The validation partition contained 133 seen-content and 479 unseen-content utterances.

This classification is an analysis of transcript-content overlap, not evidence of intentional leakage. The original split-generation procedure could not be fully reconstructed from preserved records, so the audit is reported as a post-hoc characterization rather than a reconstruction of the original splitting intent.

2.4 Whisper Small and Full-Model Fine-Tuning

The base architecture was Whisper Small (approximately 241.7 million parameters), a multilingual encoder–decoder model (Radford et al., 2023). The reported adaptation used full-model fine-tuning; no parameter-efficient adaptation method is included in the reported experimental comparison. Training ran for three epochs with a learning rate of 1×10^{-5}, AdamW (PyTorch fused implementation), a linear learning-rate schedule, warm-up ratio 0.1, and maximum gradient norm 1.0. Per-device training and evaluation batch sizes were both 1, with gradient accumulation over 16 steps, giving an effective accumulated batch size of 16. Mixed precision used FP16, gradient checkpointing was disabled, the random seed was 42, checkpoints were saved every 100 steps, and logs were recorded every 25 steps. The final selected artifact was checkpoint-918 at epoch 3.

The logged training loss decreased from approximately 47.971997 to 6.260655 across optimization. Recorded validation loss was 0.867482 for epoch 1 and 0.770586 for epoch 2. No epoch-3 validation-loss value was recovered from the preserved training logs; therefore, no value is interpolated or fabricated. Figure 2(a) visualizes the recovered training-loss trajectory together with the recorded validation-loss observations; no epoch-3 validation value is interpolated.

Table 2. Fine-tuning and controlled decoding configuration.

Parameter	Value
Base model	Whisper Small (openai/whisper-small)
Fine-tuning type	Full-model fine-tuning
Final checkpoint	checkpoint-918 (epoch 3)
Epochs / global steps	3 / 918
Learning rate	1×10^{-5}
Optimizer / scheduler	AdamW (PyTorch fused) / linear
Warm-up ratio / max gradient norm	0.1 / 1.0
Per-device train / eval batch	1 / 1
Gradient accumulation	16
Mixed precision	FP16
Random seed	42
Controlled processor	openai/whisper-small
Controlled audio preparation	mono waveform; shared resampling to 16,000 Hz
Language / task	Indonesian / transcribe
Decoding	do_sample=False; num_beams=1
Repetition penalty / no-repeat n-gram	1.0 / 0
Shared generation safety cap	225 tokens

2.5 Transcript Normalization and Evaluation Metrics

WER was calculated as (S + D + I) / N, where S, D, and I denote word substitutions, deletions, and insertions and N is the number of reference words. CER was computed analogously at character level after normalization. WER is not mathematically bounded at 100%; it can exceed 100% when insertions and other word errors together exceed the number of reference words.

2.6 Controlled Zero-Shot and Fine-Tuned Evaluation

The original Whisper Small model and checkpoint-918 were re-evaluated on the identical locked set of 613 test utterances. Both systems used the same processor, the same audio-preparation path, 16-kHz resampling, Indonesian language configuration, transcription task, deterministic one-beam decoding, and the same 225-token generation safety cap. This controlled protocol is used as the primary baseline comparison in the revised manuscript.

Historical results are retained only as reproducibility references because the exact historical fine-tuned model.generate() invocation could not be recovered from the preserved code and logs. This separation prevents historical and newly controlled metrics from being silently mixed.

To isolate the effect of supervised full-model adaptation as far as permitted by the preserved experimental artifacts, the zero-shot and fine-tuned systems were compared under identical test and decoding conditions. No additional model-size or parameter-efficient adaptation baseline was introduced in this revision because the preserved artifacts were insufficient to reconstruct an additional training condition on the exact original train–validation–test partition. Introducing a newly reconstructed split would confound the adaptation strategy with differences in data partitioning. The additional evidence for the effect of fine-tuning is therefore derived from the controlled overall comparison, the seen–unseen content analysis, and the word-error decomposition, all of which retain the original locked test evaluation.

3. Results and Discussion

3.1 Controlled Overall Performance

Under the controlled protocol, zero-shot Whisper Small produced an overall WER of 120.8384% and CER of 61.9955%. The fine-tuned checkpoint reduced WER to 51.7774% and CER to 16.1916%. The absolute WER reduction was 69.0610 percentage points, corresponding to a 57.15% relative reduction. The absolute CER reduction was 45.8039 percentage points (73.88% relative). These values are the primary performance numbers used throughout the revised manuscript. The controlled WER and CER comparisons are summarized visually in Figure 2(b) and Figure 2(c), respectively.

Table 3. Controlled zero-shot and fine-tuned performance on the locked final test.

Subset	n	Zero-shot WER	Fine-tuned WER	WER abs. reduction	Zero-shot CER	Fine-tuned CER
Overall	613	120.8384%	51.7774%	69.0610 pp	61.9955%	16.1916%
Seen-content	136	94.5122%	4.1159%	90.3963 pp	45.7344%	1.8763%
Unseen-content	477	125.8837%	60.9115%	64.9722 pp	64.5298%	18.4227%

3.2 Seen-Content and Unseen-Content Performance

The post-hoc content audit exposed a large difference in recognition difficulty. On the 136 seen-content utterances, fine-tuning reduced WER from 94.5122% to 4.1159% and CER from 45.7344% to 1.8763%. On the larger 477-utterance unseen-content subset, zero-shot WER was 125.8837% and fine-tuned WER was 60.9115%; CER decreased from 64.5298% to 18.4227%. Thus, fine-tuning improved both subsets, but unseen-content recognition remained substantially more difficult.

The seen/unseen gap is important for interpretation. Aggregate WER alone would combine a small, very easy seen-content subset with a much larger and more difficult unseen-content subset. For this corpus, the unseen-content result is therefore more informative about transcript-content generalization, while still not representing speaker-independent generalization because all three speakers were observed during training.

3.3 Controlled Word-Error Composition

The controlled word-level decomposition shows that substitutions were the dominant error category for both systems. Zero-shot evaluation produced 3,092 substitutions, 668 deletions, and 1,169 insertions, for 4,929 total word errors. After fine-tuning, the counts decreased to 1,538 substitutions, 359 deletions, and 215 insertions, for 2,112 total word errors. Fine-tuning therefore reduced all three error classes, with a particularly large drop in insertion errors. Figure 2(d) provides the corresponding visual comparison of the three word-error categories.

Taken together, the controlled results provide three complementary views of the adaptation effect. First, the overall comparison shows a reduction in WER from 120.8384% to 51.7774%. Second, the improvement remains substantial on the larger unseen-content subset, where WER decreases from 125.8837% to 60.9115%, indicating that the aggregate improvement is not

explained solely by test utterances whose normalized transcript content occurs in training. Third, fine-tuning reduces substitutions, deletions, and insertions rather than improving only one error category. These analyses are complementary controlled evaluations of the same fine-tuned system and should not be interpreted as a separate architectural ablation study.

Table 4. Controlled word-error composition.

Model	Subset	n	Substitutions	Deletions	Insertions	Hits	Total word errors
Zero-shot	Overall	613	3,092	668	1,169	319	4,929
Zero-shot	Seen	136	461	129	30	66	620
Zero-shot	Unseen	477	2,631	539	1,139	253	4,309
Fine-tuned	Overall	613	1,538	359	215	2,182	2,112
Fine-tuned	Seen	136	16	6	5	634	27
Fine-tuned	Unseen	477	1,522	353	210	1,548	2,085

3.4 Zero-Shot Long-Generation Behavior

A targeted audit was conducted on the 10 zero-shot test outputs that reached the shared 225-token generation safety cap. None of the 10 cases emitted an end-of-sequence token before a higher diagnostic ceiling, and 9 of the 10 showed strong repetitive degeneration. Increasing the ceiling therefore extended repetitive output rather than revealing a plausibly complete transcription. The 225-token limit was retained symmetrically for both systems as a safety mechanism, and the affected cases were not removed from evaluation. This behavior helps explain the unusually high zero-shot insertion burden and why zero-shot WER can exceed 100%.

3.5 Historical versus Controlled Reproducibility

The controlled re-evaluation reproduced the central scientific direction of the historical experiment. Historical fine-tuned overall WER was 52.5374%, whereas the controlled value was 51.7774%, an absolute deviation of 0.7600 percentage points. For unseen-content utterances, historical and controlled fine-tuned WERs were 61.9048% and 60.9115%, respectively, differing by 0.9933 percentage points. The fine-tuned model is therefore practically stable across the historical and controlled evaluations.

Zero-shot results were less stable across protocols: historical overall WER was 136.4550% compared with 120.8384% under the controlled protocol. This larger deviation is consistent with the separately documented long-generation behavior. For this reason, the controlled comparison is the primary reviewer-revision baseline, while historical values remain explicitly labeled as reproducibility references.

Table 5. Historical versus controlled reproducibility.

Subset	Historical zero-shot WER	Controlled zero-shot WER	Historical fine-tuned WER	Controlled fine-tuned WER
Overall	136.4550%	120.8384%	52.5374%	51.7774%
Seen	95.1220%	94.5122%	3.6585%	4.1159%
Unseen	144.3763%	125.8837%	61.9048%	60.9115%

3.6 Effect of Full-Model Fine-Tuning

The controlled comparison shows that full-model adaptation substantially improves Whisper Small on the processed Makassar corpus. Because the zero-shot and fine-tuned systems were evaluated on the same locked test items using matched preprocessing, normalization, language/task configuration, and deterministic decoding conditions, the observed difference is not attributable to unequal inference settings. Overall WER decreased from 120.8384% to 51.7774%, while CER decreased from 61.9955% to 16.1916%. These results demonstrate a substantial adaptation benefit within the available speaker-dependent corpus, although the remaining error rate indicates that Makassar recognition is far from solved.

The direction of this improvement is consistent with previous low-resource Whisper research, but the experimental emphasis differs. Liu et al. (2024) investigated multiple fine-tuning

strategies across seven low-resource languages and showed that different adaptation strategies provide different performance and efficiency characteristics. Parameter-efficient adaptation has also been investigated by Ghimire et al. (2024), who applied adapter-based fine-tuning to Whisper in a low-resource setting. Regional-language research has similarly demonstrated the potential of Whisper adaptation, including recent work on Mandar (Jafar et al., 2025). These studies provide important evidence for the general effectiveness of target-language adaptation. The present study does not attempt to establish a superior adaptation strategy; instead, it examines how the benefit of full-model fine-tuning should be interpreted when transcript-content overlap, restricted speaker diversity, error composition, and evaluation reproducibility are considered explicitly. More recent work combining Whisper with LoRA and hyperparameter optimization likewise illustrates the active development of parameter-efficient low-resource adaptation (Wang et al., 2025). In parallel, the Makassar benchmark reported by Azis et al. (2026) highlights the need for language-specific evaluation, although its spontaneous-speech benchmark and multi-model comparison differ from the controlled fine-tuning design used here. Related Hawaiian work has also shown that language-specific text resources can complement Whisper through external language-model rescoring, illustrating another adaptation path for underrepresented languages (Chaparala et al., 2024).

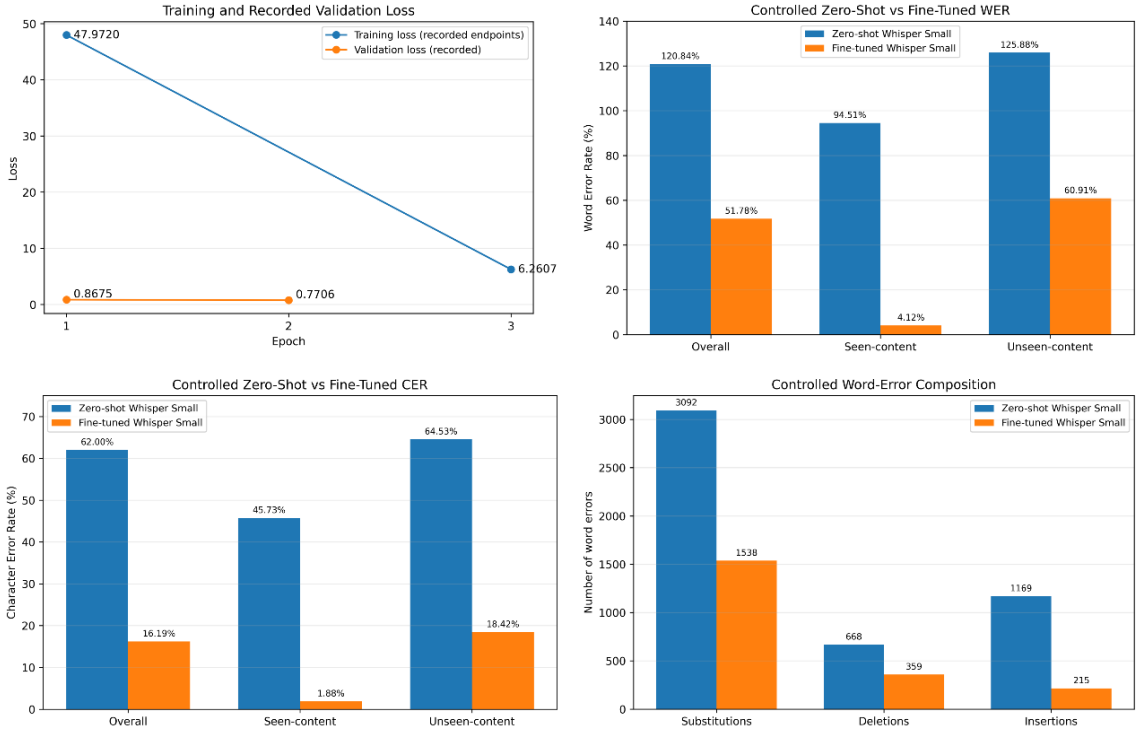


Figure 2. Key experimental results: (a) training and recorded validation loss (no epoch-3 validation loss was logged); (b) controlled WER comparison; (c) controlled CER comparison; and (d) controlled word-error composition (substitutions, deletions, and insertions).

This distinction is important because a reduction in aggregate WER alone does not establish the scope of generalization. In the present experiment, the improvement remains substantial on the 477 unseen-content utterances, where WER decreases from 125.8837% to 60.9115%. The adaptation effect therefore extends beyond normalized transcript content represented during training. However, because the same three speakers occur across the training, validation, and test partitions, this result represents generalization to unseen transcript content under a speaker-dependent setting rather than generalization to unseen speakers.

3.7 Transcript-Content Generalization and Evaluation Design

The large performance difference between seen and unseen transcript content demonstrates why aggregate error rates require careful interpretation in small-corpus ASR. Fine-tuned WER was 4.1159% on the 136 seen-content utterances but 60.9115% on the 477 unseen-content utterances. If only the overall WER of 51.7774% were reported, this substantial difference in recognition difficulty would remain hidden.

At the same time, the unseen-content result provides an important robustness check on the main finding. The controlled zero-shot WER for these 477 utterances was 125.8837%, compared with 60.9115% after fine-tuning. Thus, the overall adaptation benefit cannot be explained solely by the 136 test utterances whose normalized transcript content also occurs in training. Nevertheless, the remaining unseen-content WER shows that recognition of previously unobserved transcript content remains substantially more difficult than recognition of seen content.

This content-based analysis must not be conflated with speaker-independent evaluation. Lukas, Markus, and Matius occur in all three data partitions; therefore, both the seen- and unseen-content results remain speaker-dependent. A stronger test of generalization would require speaker-disjoint evaluation, while a deliberately transcript-disjoint benchmark would provide a complementary test of content generalization. These designs should be prioritized in future Makassar ASR studies.

3.8 Interpreting WER above 100%

The zero-shot WER above 100% is valid rather than a calculation anomaly. WER counts substitutions, deletions, and insertions relative to the number of reference words and has no upper bound at 100%. In the controlled zero-shot condition, 1,169 insertion errors contributed heavily to the 4,929 total word errors. The targeted long-generation audit further showed that some ceiling-hit outputs became repetitive and failed to terminate normally. Together, these observations provide a concrete explanation for the high zero-shot WER. This diagnostic finding reinforces the broader evaluation argument of the study: ASR performance should be interpreted not only from the final aggregate WER but also from the error composition and decoding behavior that produce it.

3.9 Choice of Whisper Small

Whisper Small was selected as a practical experimental base model that balanced model capacity with the computational cost of full-model fine-tuning and repeated evaluation in the available environment. The present study therefore evaluates one specific Whisper configuration rather than claiming that Small is universally optimal for Makassar ASR. A systematic comparison with Tiny, Base, Medium, Large, and alternative multilingual speech models remains appropriate future work. This distinction is important because model-size comparisons should be established experimentally rather than inferred from the current single-architecture study. Accordingly, the contribution of this study should not be interpreted as a model-selection result. Whisper Small serves as the controlled experimental architecture through which the effects and evaluation boundaries of full-model adaptation are examined. Comparisons with other Whisper sizes, parameter-efficient adaptation methods, or alternative multilingual ASR architectures require separate experiments under an identical data partition and evaluation protocol.

3.10 Limitations

The study has several limitations. First, the corpus contains only three speakers, and the same speakers occur in training, validation, and test. The reported evaluation is therefore speaker-dependent and cannot establish performance on unseen speakers or the broader Makassar-speaking population.

Second, 136 of 613 test utterances (22.19%) share normalized transcript content with training. The separate unseen-content analysis reduces the risk of hiding this effect, but the original split was not designed as a transcript-disjoint benchmark.

Third, the exact historical fine-tuned model.generate() invocation could not be recovered from preserved code and logs. This limitation motivated the controlled re-evaluation and the explicit separation of historical from controlled results.

Fourth, zero-shot Whisper Small exhibited repetitive long-generation behavior in 10 ceiling-hit cases; 9 showed strong repetition and none emitted EOS before the higher diagnostic ceiling. This behavior should be considered when interpreting the zero-shot insertion burden.

Fifth, validation loss was recovered only for epochs 1 and 2. No epoch-3 validation-loss value was logged in the preserved record, so the loss curve cannot be presented as a complete three-epoch validation trajectory.

Sixth, the experimental comparison is limited to zero-shot and full-model fine-tuned Whisper Small. The study therefore demonstrates the benefit of adaptation relative to the pretrained baseline but does not determine whether full-model fine-tuning is preferable to parameter-efficient adaptation, alternative Whisper sizes, or other ASR architectures. An additional training baseline was not reconstructed during the revision because the preserved artifacts did not support reproducing such a condition on the exact original partition. Future comparative experiments should retain a common train–validation–test partition and decoding protocol so that performance differences can be attributed to the adaptation strategy rather than to changes in data composition or inference conditions. Future controlled extensions may also evaluate data-augmentation strategies such as SpecAugment, which masks time and frequency regions in acoustic features and has been widely used in ASR (Park et al., 2019).

Finally, the source recordings and underlying Makassar Bible text are third-party copyrighted materials. SABDA identifies the Makassar Bible text as © 1999 Lembaga Alkitab Indonesia (LAI) and the corresponding audio recording as ℗ 2012 Hosanna. Accordingly, this study does not claim redistribution rights over the original audio or source text, and the source materials are not redistributed with the research artifacts.

4. Conclusions

This study shows that full-model fine-tuning substantially improves Whisper Small for the available low-resource Makassar ASR corpus when the baseline and adapted model are compared under a shared controlled protocol. On the locked 613-utterance test set, overall WER decreased from 120.8384% to 51.7774%, while CER decreased from 61.9955% to 16.1916%. The improvement also persisted on the 477 unseen-content utterances, where WER decreased from 125.8837% to 60.9115%, an absolute reduction of 64.9722 percentage points.

These findings demonstrate the value of full-model adaptation for the available Makassar corpus while also illustrating the importance of evaluation design in very low-resource ASR. The substantial difference between seen- and unseen-content performance shows why aggregate error rates should be interpreted together with transcript-content visibility, while the shared-speaker design prevents the unseen-content result from being interpreted as unseen-speaker generalization. The reproducibility and decoding audits further show that evaluation protocol and generation behavior can materially affect the interpretation of baseline performance. Future work should prioritize speaker-disjoint and transcript-disjoint evaluation, broader speaker diversity, controlled comparisons with parameter-efficient adaptation and alternative model sizes, and decoding strategies that reduce repetitive generation.

Data Availability and Research Transparency

The source Makassar-language recordings used to construct the experimental corpus were obtained from the publicly accessible SABDA Audio platform (https://audio.sabda.org/). SABDA's copyright information identifies the Makassar Bible text as © 1999 Lembaga Alkitab Indonesia (LAI) and the corresponding audio recording as ℗ 2012 Hosanna. Because the underlying recordings and source text are third-party copyrighted materials, the original audio files and source text are not redistributed by the authors. Access to the source material should be obtained directly from SABDA and remains subject to the applicable rights and usage conditions. Derived manifests/metadata, controlled evaluation tables, and evaluation scripts used for the

reported analysis are available from the corresponding author upon reasonable request. The model checkpoint is not included in the shared research artifacts.

Data Availability and Research Transparency

The source Makassar-language recordings used to construct the experimental corpus were obtained from the publicly accessible SABDA Audio platform (https://audio.sabda.org/). SABDA's copyright information identifies the Makassar Bible text as © 1999 Lembaga Alkitab Indonesia (LAI) and the corresponding audio recording as ℗ 2012 Hosanna. Because the underlying recordings and source text are third-party copyrighted materials, the original audio files and source text are not redistributed by the authors. Access to the source material should be obtained directly from SABDA and remains subject to the applicable rights and usage conditions. Derived manifests/metadata, controlled evaluation tables, and evaluation scripts used for the reported analysis are available from the corresponding author upon reasonable request. The model checkpoint is not included in the shared research artifacts.

Acknowledgements

The authors acknowledge the academic institutions and collaborators who supported the Makassar ASR research and the computational experimentation. The authors also acknowledge SABDA Audio as the public access platform for the source recordings and Lembaga Alkitab Indonesia (LAI) and Hosanna as the rights holders identified by SABDA for the Makassar Bible text and audio recording, respectively.

References

Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., Alharbi, S., Alturki, S., Alshehri, F., & Almojil, M. (2021). Automatic speech recognition: Systematic literature review. IEEE Access, 9, 131858–131876. https://doi.org/10.1109/ACCESS.2021.3112535

Slam, W., Li, Y., & Urouvas, N. (2023). Frontier research on low-resource speech recognition technology. Sensors, 23(22), 9096. https://doi.org/10.3390/s23229096

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 28492–28518). PMLR.

Liu, Y., Yang, X., & Qu, D. (2024). Exploration of Whisper fine-tuning strategies for low-resource ASR. EURASIP Journal on Audio, Speech, and Music Processing, 2024, 29. https://doi.org/10.1186/s13636-024-00349-3

Pratama, R. S. A., & Amrullah, A. (2024). Analysis of Whisper automatic speech recognition performance on low resource language. Jurnal Pilar Nusa Mandiri, 20(1), 1–8. https://doi.org/10.33480/pilar.v20i1.4633

Jafar, J., Tb, M. A. W., Aziz, F., Iriany, R., & Nasir, N. (2025). Fine-tuning Whisper model for Mandar speech recognition: Approach and performance evaluation. Journal of Applied Engineering and Technological Science, 7(1), 261–273. https://doi.org/10.37385/jaets.v7i1.7170

Polat, H., Turan, A. K., Koçak, C., & Ulaş, H. B. (2024). Implementation of a Whisper architecture-based Turkish automatic speech recognition (ASR) system and evaluation of the effect of fine-tuning with a Low-Rank Adaptation (LoRA) adapter on its performance. Electronics, 13(21), 4227. https://doi.org/10.3390/electronics13214227

Ghimire, R. R., Poudyal, P., & Bal, B. K. (2024). Improving on the limitations of the ASR model in low-resourced environments using parameter-efficient fine-tuning. In Proceedings of the 21st International Conference on Natural Language Processing (ICON) (pp. 408–415).

Chaparala, K. D., Zarrella, G., Fischer, B. T., Kimura, L., & Jones, O. P. (2024). Mai Hoʻomāuna i ka ʻAi: Language models improve automatic speech recognition in Hawaiian. In Proceedings of the 4th NeurIPS Efficient Natural Language and Speech Processing Workshop (Vol. 262, pp. 576–583). PMLR.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., & Auli, M. (2024). Scaling speech technology to 1,000+ languages. Journal of Machine Learning Research, 25(97), 1–52.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., & Weber, G. (2020). Common Voice: A massively-multilingual speech corpus. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 4218–4222). European Language Resources Association.

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., & Bapna, A. (2023). FLEURS: Few-shot learning evaluation of universal representations of speech. In 2022 IEEE Spoken Language Technology Workshop (SLT) (pp. 798–805). IEEE. https://doi.org/10.1109/SLT54892.2023.10023141

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449–12460.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., & Auli, M. (2022). XLS-R: Self-supervised cross-lingual speech representation learning at scale. In Interspeech 2022 (pp. 2278–2282). https://doi.org/10.21437/Interspeech.2022-143

Sakti, S., & Titalim, B. A. (2023). Leveraging the multilingual Indonesian ethnic languages dataset in self-supervised models for low-resource ASR task. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 1–8). IEEE. https://doi.org/10.1109/ASRU57964.2023.10389730

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In Proceedings of the 36th International Conference on Machine Learning (Vol. 97, pp. 2790–2799). PMLR.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A simple augmentation method for automatic speech recognition. In Interspeech 2019 (pp. 2613–2617). https://doi.org/10.21437/Interspeech.2019-2680

Wang, H., Bin, J., Gou, C., Yang, L., Hou, B., & Qin, M. (2025). Low-resource speech recognition by fine-tuning Whisper with Optuna-LoRA. Applied Sciences, 15(24), 13090. https://doi.org/10.3390/app152413090

Azis, H., Fajriansyah M, M. F., Darwis, H., Purnawansyah, Hasanuddin, T., & Sugiarti. (2026). Development of a low-resource automatic speech recognition system for the Makassar dialect. In 20th International Conference on Ubiquitous Information Management and Communication (IMCOM) (pp. 1–8). IEEE. https://doi.org/10.1109/IMCOM69009.2026.11360893