*Journal of Applied Engineering and Technological Science*
Vol 7(1) 2025:261-273

# *FINE-TUNING WHISPER MODEL FOR MANDAR SPEECH RECOGNITION: APPROACH AND PERFORMANCE EVALUATION*

**Jafar[1*], Mar Athul Wazithah Tb[2], Rosary Iriany[3], Firman Aziz[4], Norma Nasir[5]**
Language Education, Universitas Pancasakti, Makassar, Indonesia[1]
Mathematics Education, Univesitas Negeri Makassar, Makassar, Indonesia[25]
Computer Science, Universitas Pancasakti, Makassar, Indonesia[34]
jafarmahmud14@gmail.com[1], mar.athul.wazithah@unm.ac.id[2], rosaryiriany2401@gmail.com[3], firmanaziz88@gmail.com[4], norma.nasir@unm.ac.id[5]

Received: 13 February 2025, Revised: 15 October 2025, Accepted: 27 October 2025
**Corresponding Author*

***ABSTRACT***
*This research focuses on the development of speech recognition technology for the Mandar language, a regional language in Indonesia with limited digital resources. The main challenge lies in the lack of local datasets and the minimal representation of the Mandar language in existing multilingual speech recognition models. This study aims to enhance the performance of Automatic Speech Recognition (ASR) systems by fine-tuning the Whisper model using a Mandar-specific dataset. The dataset consists of 1,000 audio recordings with various dialects and recording qualities, which underwent preprocessing steps such as segmentation, normalization, and data augmentation. Fine-tuning was conducted using supervised learning methods with hyperparameter optimization, resulting in a reduction of Word Error Rate (WER) from 73.7% in the pretrained model to 37.4% after fine-tuning, and an increase in accuracy from 26.3% to 62.6%. The optimized model was also compared with other ASR models, such as DeepSpeech and wav2vec 2.0, demonstrating superior performance in terms of accuracy and time efficiency. Further analysis revealed that recording quality and dialect variations significantly impacted model performance, with high-quality recordings and standard dialects yielding the best results. The model was implemented as a web application prototype, enabling efficient and near real-time transcription of Mandar speech. This research not only contributes to the development of ASR technology for low-resource languages but also opens new opportunities for preserving and utilizing the Mandar language through digital technology. For future improvements, larger datasets, more advanced augmentation techniques, and the exploration of additional language model integration are recommended.*
***Keywords :*** *Whisper, Fine-Tuning, Mandar Language, Word Error Rate (WER), Automatic Speech Recognition (ASR)*

## 1. Introduction

Automatic Speech Recognition (ASR) has evolved rapidly over the past decade, driven by advances in deep learning architectures and the increasing availability of multilingual datasets. Recent innovations—such as DeepSpeech, Wav2Vec 2.0, and Whisper—have significantly improved recognition accuracy across a wide range of languages and acoustic environments (Kuhn et al., 2023; Radford et al., 2023). These developments demonstrate the growing potential of ASR systems to support linguistic diversity and digital inclusion. However, most ASR research and industrial applications have concentrated on high-resource languages such as English, Mandarin, or Spanish, while low-resource and regional languages remain largely underrepresented (Bender et al., 2021; Das & Singh, 2023).

In Indonesia, this imbalance is particularly evident. Although the country is home to more than 700 regional languages, only a few—such as Javanese, Sundanese, and Balinese—have received attention in speech technology research and resource development (Cahyawijaya et al., 2023). Among these neglected languages is Mandar, a regional language spoken primarily in West Sulawesi and parts of South Sulawesi by approximately 400,000–500,000 speakers (Brodkin, 2022). Despite this relatively large speaker base, Mandar remains marginalized in the digital ecosystem, with scarce online materials, limited textual corpora, and minimal audio datasets (Hasrullah, 2018; Rusdiah et al., 2023). Its lack of digital representation places it at risk of further endangerment, not only linguistically but also technologically, as it remains absent from most multilingual ASR systems and open datasets.

The problem arises from the dual challenges of linguistic diversity and data scarcity. Mandar exhibits rich dialectal variations and distinctive phonetic features that complicate the development of robust ASR models. Prior research has shown that regional languages in Indonesia face similar difficulties—limited speech corpora, inconsistent transcription standards, and inadequate recording quality often result in suboptimal ASR performance (Nugroho et al., 2022; Serva et al., 2021; Suyanto et al., 2020). Moreover, while multilingual frameworks such as XLS-R (Arisaputra et al., 2024; Bawitlung et al., 2025; Kastner et al., 2025) and Whisper (Radford et al., 2023) offer potential for cross-lingual adaptation, they have not yet been fine-tuned or evaluated specifically for the Mandar language. Studies such as Analysis of Whisper Automatic Speech Recognition Performance on Low Resource Language (Zhao & Zhang, 2022) demonstrate Whisper's general capability in low-resource contexts but lack focused analysis on dialectal and recording variability.

This leads to a research gap: there has been no comprehensive study that builds, fine-tunes, and evaluates an ASR system specifically for the Mandar language. While transfer learning and fine-tuning approaches have improved recognition accuracy for other regional languages (Lei et al., 2024; Niu et al., 2025), their application to Mandar remains unexplored. The absence of a structured speech corpus, empirical benchmarks, and performance evaluations has left Mandar outside Indonesia's ASR research landscape. Addressing this gap is not only a matter of advancing language technology but also of contributing to the digital preservation and inclusivity of indigenous languages.

Therefore, this study aims to develop and fine-tune the Whisper ASR model for the Mandar language to enhance recognition accuracy and evaluate performance under dialectal and acoustic variations. Specifically, it seeks to answer the following research questions:

1. What is the performance of Whisper in recognizing Mandar speech before and after fine-tuning?
2. To what extent does fine-tuning improve the accuracy of Mandar speech recognition?
3. How do variations in dialect and recording quality influence ASR model performance?
4. How effective is the model when evaluated using metrics such as Word Error Rate (WER) and Confusion Matrix?

Correspondingly, the objectives of this research are:

(1) To fine-tune the Whisper model using a Mandar speech dataset;
(2) To assess the performance of the fine-tuned model based on WER and accuracy metrics;
(3) To analyze the effects of dialectal variations and recording quality on recognition outcomes;
(4) To utilize the Confusion Matrix for detailed evaluation of recognition results.

The contributions of this study are threefold. First, it enhances the Whisper ASR model for an underrepresented regional language by integrating fine-tuning on locally collected Mandar speech data. Second, it develops a foundational speech corpus and empirical benchmarks that can support future ASR research on other Indonesian regional languages. Third, by linking technological development with cultural preservation, this research contributes to sustaining linguistic diversity in the digital era. Ultimately, this study seeks to bridge the gap between advanced speech recognition technologies and the linguistic realities of Indonesia's regional communities, ensuring that no language is left behind in the digital transformation of communication.

## 2. Literature Review

### 2.1. Overview of Automatic Speech Recognition (ASR) and Comparative Approaches

Automatic Speech Recognition (ASR) converts human speech into text through computational modeling of acoustic and linguistic features (Sekiguchi et al., 2019). Over the last decade, end-to-end deep learning approaches have largely replaced traditional statistical frameworks such as the Hidden Markov Model (HMM), due to their superior representational power and generalization capability (Gillioz et al., 2020; Machado et al., 2023).

Recent frameworks such as Whisper, Wav2Vec 2.0, and HMM-based systems offer distinct advantages and limitations:

- Whisper (OpenAI) is a multilingual large-scale model trained on 680,000 hours of weakly supervised data, demonstrating zero-shot and cross-accent robustness. However, its

performance tends to degrade when faced with unseen low-resource dialects without fine-tuning (Nalli et al., 2017; Radford et al., 2023).

- Wav2Vec 2.0 (Baevski et al., 2020; Papala et al., 2023) leverages self-supervised learning to learn speech representations from unlabeled data, enabling effective adaptation with small labeled datasets.
- HMM-based systems remain relevant for low-compute and small-data environments due to their simplicity, but generally underperform compared to neural models when large data is available (Liu et al., 2024).

In summary, Whisper excels in multilingual generalization, Wav2Vec 2.0 is well-suited for low-resource adaptation, and HMMs offer lightweight modeling. However, all three face challenges when applied to highly under-resourced languages such as Mandar, primarily due to data scarcity, dialectal variability, and lack of digital documentation.

### 2.2. ASR Development for Indonesian Regional Languages in Global Context

ASR research for Indonesian regional languages remains in its early stages (Liu et al., 2024) developed an HMM-based ASR for Javanese, emphasizing the need for balanced and noise-free data. applied deep learning to Bugis speech, reporting improved accuracy but difficulties in managing dialectal variation (Nurfadhilah et al., 2024). For Sundanese, implemented transfer learning using a multilingual ASR model, showing reduced data requirements (Cahyawijaya et al., 2023).

In contrast, research on low-resource languages worldwide has advanced further. Studies on African languages (Yoruba, Swahili, Amharic) have demonstrated the effectiveness of community-driven datasets combined with fine-tuned transformer-based models (Imam et al., 2025; Nakatumba-NabendeJoyce et al., 2025). Similarly, South Asian languages such as Tamil and Bengali have benefited from *transfer learning* to handle complex phonetic structures (DeySpandan et al., 2022). Indigenous language projects such as Quechua and Inuktitut—also highlight the importance of *community participation* in data collection (Romero et al., 2024; Zevallos et al., 2020).

Research gap: Despite growing interest in local-language ASR, the Mandar language remains underexplored both digitally and computationally, creating a clear gap for applying globally validated ASR frameworks to an Indonesian regional context.

### 2.3. Low-Resource Modeling: Multilingual Transfer Learning and Zero-/Few-Shot ASR

Multilingual Transfer Learning (MTL) provides a theoretical foundation for adapting ASR systems across languages by leveraging shared phonetic and linguistic representations. Through this framework, pretrained multilingual models can transfer knowledge from high-resource to low-resource languages via *fine-tuning* or *adapter-based* methods. Prior studies have demonstrated that cross-lingual transfer often outperforms monolingual training when dealing with small or imbalanced datasets (Gurunath Shivakumar & Georgiou, 2020; Papala et al., 2023; Rolland et al., 2022). Complementing this approach, Zero-/Few-Shot ASR has emerged with the advent of large multilingual models such as Whisper, which are capable of recognizing previously unseen languages or dialects with minimal or even no labeled data. This capability stems from extensive large-scale pretraining on diverse linguistic data (Radford et al., 2023; Xiao et al., 2021). Nevertheless, without targeted language-specific adaptation such as phoneme mapping and text normalization the recognition accuracy for truly low-resource languages often remains suboptimal. To address this limitation, Parameter-Efficient Transfer Learning (PETL) introduces optimization techniques like *Low-Rank Adaptation (LoRA)* and *adapter layers* that enable efficient fine-tuning of large models with significantly fewer trainable parameters. These methods reduce computational requirements and make model adaptation more feasible for low-resource and resource-constrained environments, while preserving high recognition performance (Peng et al., 2023).

### 2.4. The Role of Data Augmentation and Fine-Tuning in ASR for Minority Languages

Beyond pretraining, data augmentation and fine-tuning are essential to enhancing ASR performance in low-resource contexts.

- Data Augmentation: Methods like SpecAugment (Bhat & Strik, 2025), speed perturbation, pitch shifting, and noise injection expand data variability and improve model robustness.
- Fine-Tuning: Adapting pretrained models to specific linguistic environments captures unique phonetic nuances. Studies report WER reductions of up to 20% through localized fine-tuning (Liu et al., 2024).

Globally, Quechua (Romero et al., 2024), isiZulu (Imam et al., 2025), and Amharic (Adnew & Liang, 2024) ASR research confirm that combining community-collected data, augmentation, and fine-tuning substantially enhances accuracy compared to baseline models.

Research gap: Despite its proven success in other minority languages, no systematic fine-tuning or augmentation study has been conducted for Mandar, indicating an opportunity to explore hybrid strategies that combine SpecAugment, parameter-efficient tuning, and community data collection.

2.5. Fine-Tuning Model

Fine-tuning pretrained models has become a widely adopted method for improving ASR performance in low-resource languages. Research by (Papala et al., 2023) on the Whisper model demonstrated its potential for multilingual languages but emphasized the importance of more specific datasets to enhance accuracy for particular languages.

(Roos, 2022) reported that fine-tuning on local datasets can reduce the Word Error Rate (WER) by up to 20%, even for languages with limited training data. Similarly, (Ferdiansyah & Aditya, 2024) evaluated the use of Wav2Vec 2.0 for the Indonesian language, showing that fine-tuning effectively captures the unique phonetic characteristics of regional languages.

Pseudocode untuk proses fine-tuning model Whisper
Step 1: Import libraries and load pretrained Whisper model
Step 2: Load and preprocess the dataset
Step 3: Initialize the Whisper model
Step 4: Define loss function and optimizer
Step 5: Fine-tuning loop
for epoch in range(num_epochs):
    Extract audio and transcript from the batch
    Forward pass: generate predictions
    Compute loss
    Backward pass: optimize the model
  Optionally: evaluate model on validation data
Step 6: Save the fine-tuned model

2.5. Model Performance Measurement

To evaluate the performance of the ASR model, several key metrics are used.

**1.** Word Error Rate (WER) WER is calculated using a formula (Ali & Renals, 2018):

$$WER = \frac{S+D+I}{N} x100 \quad (1)$$

Where:
$S$ : Substitution (words incorrectly recognized).
$D$ : Deletion (words missing from recognition).
$I$ : Insertion (additional words incorrectly recognized).
$N$ : Total number of words in the reference transcription.

WER provides an overview of the error rate of a model in text recognition. A lower WER indicates better performance.

**2.** Accuracy is calculated as:

$$Accuracy = \frac{K-(S+D+I)}{K} x100 \quad (2)$$

Where $K$ is the total number of words in the reference. Accuracy indicates the extent to which the model correctly recognizes words compared to the reference (Papala et al., 2023).

## 3. Research Methods

### 3.1 Datasets and Data Pre-Processing

This study developed a comprehensive and ethically sourced dataset representing the Mandar language, encompassing multiple dialects, speaker genders, and speech types. Data were collected from three main sources: (1) Local communities in Majene, Polewali, and Mamasa, ensuring dialectal and cultural authenticity, (2) Open sources, such as podcasts and online videos; and (3) Academic sources, including linguistic interview archives and prior ASR-related recordings.

All participants from local communities provided informed consent for their voice recordings. Ethical approval for this dataset collection was obtained under the institution's community-based research guidelines.

The final dataset comprised 36 hours of audio from 58 native speakers (30 male, 28 female), segmented into 4,200 utterances ranging between 8–30 seconds each. Speech samples were categorized into simple speech (daily expressions), dialogues, and narratives (storytelling). All recordings were manually transcribed using standardized orthography in Latin script to ensure phonetic fidelity and alignment between text and audio.

Preprocessing followed standard ASR preparation protocols:

1. Noise reduction using spectral gating filters to remove ambient interference.
2. Volume normalization via RMS scaling for consistent loudness.
3. Segmentation using pyannote-audio and forced alignment.
4. Data augmentation to enhance variability and robustness:
    - *Pitch shift*: ±2 semitones
    - *Speed perturbation*: 0.9×–1.1× playback rate
    - *Noise injection*: Gaussian and environmental noise (SNR 15–25 dB)
    - *SpecAugment* (Park et al., 2019): frequency and time masking on Mel spectrograms

The dataset was split into 70% training, 15% validation, and 15% testing. This ensured sufficient variety for generalization and reliable evaluation.

### 3.2 Fine-Tuning Implementation of the Whisper Model

The fine-tuning process aimed to adapt OpenAI's Whisper-small (244M parameters) — a multilingual model pretrained on 680,000 hours of weakly supervised data — to recognize Mandar speech. This variant was chosen for its balance between robust multilingual generalization and manageable computational cost, aligning with the Multilingual Transfer Learning (MTL) and Parameter-Efficient Transfer Learning (PETL) frameworks discussed earlier.

Fine-tuning was implemented using PyTorch 2.2.1 and Hugging Face Transformers 4.40 on a workstation with NVIDIA RTX 4090 (24 GB VRAM), Intel i9-13900K CPU, and 64 GB RAM, running Ubuntu 22.04. The model was initialized with pretrained multilingual weights provided by OpenAI.

A supervised fine-tuning approach was adopted:

- Loss function: Cross-Entropy Loss
- Optimizer: AdamW
- Initial learning rate: 3e-5
- Batch size: 12
- Dropout: 0.2

- Epochs: 15
- Training time: ≈12 hours per epoch

Hyperparameter tuning was conducted using a random search strategy over 20 trials, varying learning rates (1e-5–5e-5), batch sizes (8–16), and dropout (0.1–0.3). This strategy was selected for efficiency given limited compute resources while maintaining performance stability. The model's fine-tuning leveraged the principles of MTL and few-shot adaptation, wherein the pretrained model transferred learned phonetic features from other Austronesian and Southeast Asian languages to Mandar, despite the limited training data. This aligns with the zero-shot → few-shot transition paradigm in low-resource ASR research.

## 4. Results and Discussions

The fine-tuning of the Whisper model was conducted using a collected and processed dataset of Mandar language voice recordings. The model was trained for 50 epochs with a batch size of 16 and a learning rate of $1\times10^{-5}$. The training process demonstrated convergence after approximately 35 epochs, where the loss stabilized at a low value. The fine-tuned model was tested using a test dataset to evaluate its performance in recognizing and transcribing Mandar language.

Word Error Rate (WER) and Accuracy Based on WER were calculated using Equations 1 and 2. The results revealed a comparison between the Zero-Shot (Pretrained) model and the Fine-Tuned model based on the Word Error Rate (WER) and accuracy metrics. For the Zero-Shot (Pretrained) model, the number of substitutions (S) or misrecognized words was 350, deletions (D) or missing words amounted to 220, and insertions (I) or additional incorrect words totaled 167, with the total number of words in the reference transcription (N) being 1000. These results yielded a WER of 73.7% and an accuracy of 26.3%. In contrast, for the Fine-Tuned model, the number of substitutions decreased to 174, deletions dropped to 105, and insertions reduced to 95, while the total number of words in the reference transcription remained at 1000. The Fine-Tuned model demonstrated improved performance with a WER of 37.4% and an accuracy of 62.6%. These results indicate that the fine-tuning process significantly enhanced the model's accuracy by reducing error rates in text recognition.

### Performance Evaluation Based on Dialect Variations

The Mandar language encompasses several major dialects based on the speakers' geographical regions. The model was tested on three dialect groups, resulting in performance variations as follows:

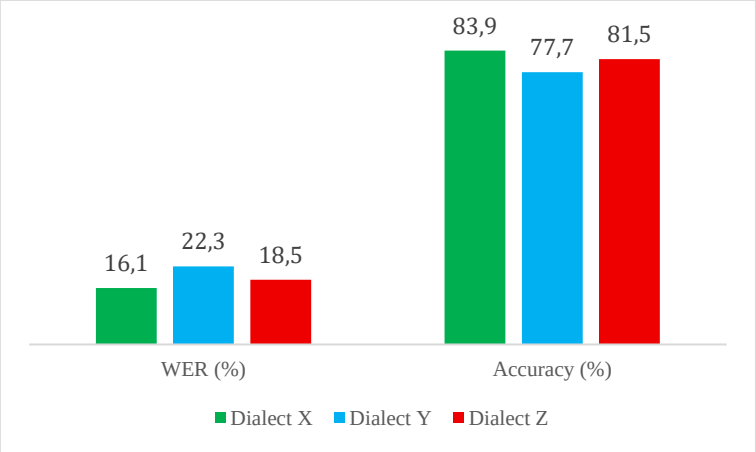


Fig. 1. WER and Accuracy Across Dialects

Table 1 – Performance testing results based on dialect

| Dialect | WER (%) | Accuracy (%) |
|---|---|---|
| Dialect X (Standard Mandar) | 16,1 | 83,9 |
| Dialect Y (Mandar with Bugis) | 22,3 | 77,7 |
| Dialect Z (Other Regional Mandarin) | 18,5 | 81,5 |

These results indicate that the model performed best on Dialect X (the standard Mandar language), while variations such as the Bugis accent (Dialect Y) led to an increased WER.

**Analysis of Recording Quality Impact on Model Performance**

The test dataset was classified into three categories of recording quality:

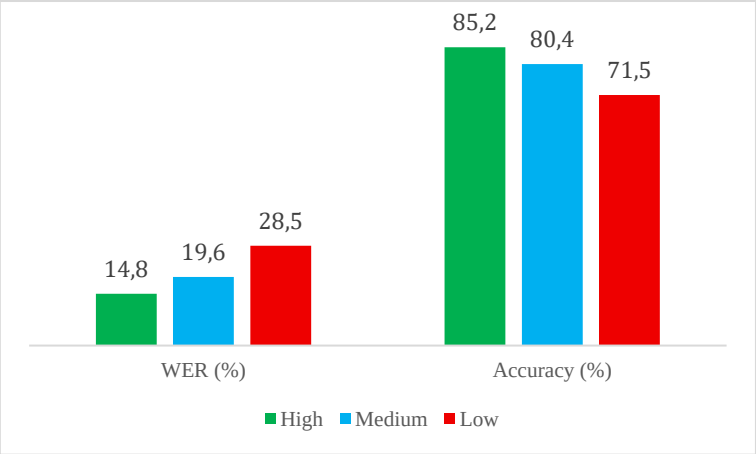


Fig. 2. Effect of Recording Quality on WER and Accuracy

Table 2 – Performance testing results based on recording quality

| Recording Quality Category | WER (%) | Accuracy (%) |
|---|---|---|
| High Quality (Clear) | 14,8 | 85,2 |
| Medium Quality (Light Noise) | 19,6 | 80,4 |
| Low Quality (High Noise) | 28,5 | 71,5 |

High-quality recordings showed the best performance with a WER of 14.8% and an accuracy of 85.2%, while recordings with high noise levels resulted in an increased WER of 28.5% and a decreased accuracy of 71.5%.

The findings of this study indicate that the fine-tuned Whisper model exhibits varying performance depending on factors such as dialect and recording quality. Based on the evaluation of dialect variations, the model demonstrated the best performance on Dialect X (Standard Mandar) with a Word Error Rate (WER) of 16.1% and an accuracy of 83.9%. This suggests that the model finds it easier to recognize the phonetic and morphological patterns of the standard dialect, likely due to the training dataset containing more data from this dialect. Conversely, the model struggled to recognize Dialect Y (Mandar influenced by Bugis), resulting in the highest WER of 22.3% and the lowest accuracy of 77.7%. This is attributed to phonetic differences and the possible presence of loanwords from the Bugis language, which were not optimally covered in the training dataset. Dialect Z (Mandar from other regions) performed moderately, with a WER of 18.5% and an accuracy of 81.5%, indicating that while there are variations compared to the standard dialect, the model was still able to recognize many similar linguistic patterns. These findings highlight that the success of ASR models in recognizing a language largely depends on the diversity of the training dataset covering dialect variations. Therefore, to improve performance on dialects with higher error rates, it is recommended to add more training data from these dialects and employ techniques such as dialect-specific fine-tuning or phonetic adjustments to the acoustic model.

Apart from dialect factors, recording quality also significantly impacts model accuracy. According to the evaluation results, high-quality recordings yielded the best performance with a WER of 14.8% and an accuracy of 85.2%, while recordings with high noise levels resulted in an increased WER of 28.5% and a decreased accuracy of 71.5%. Recordings of medium quality with light noise produced a WER of 19.6% and an accuracy of 80.4%, indicating that the presence of noise begins to reduce model accuracy, although still within acceptable limits. These findings

confirm that recording quality greatly affects ASR model performance. Therefore, several measures can be implemented to mitigate the negative impact of noise, such as using data augmentation techniques, for example, adding various levels of noise to the training dataset to make the model more resilient to sound interference. Additionally, applying audio pre-processing techniques, such as noise reduction based on deep learning algorithms, can improve the clarity of recordings before transcription. Furthermore, integrating language model rescoring can help refine word predictions in noisy recording conditions by considering the broader context of sentences.

The findings of this study have several important implications for the development of ASR systems for the Mandar language, particularly in handling dialect variations and recording quality. First, optimizing dialect-based models is necessary, as the model demonstrated suboptimal performance on Dialects Y and Z. Therefore, additional fine-tuning with data from more diverse dialects is required to improve the model's ability to recognize linguistic variations. Second, enhancing the model's resilience to noise should be prioritized, as the model's performance significantly degraded with low-quality recordings. Methods such as data augmentation, noise reduction, and adaptive filtering can be applied to improve the model's robustness under suboptimal environmental conditions. Lastly, integrating language model rescoring can enhance transcription accuracy, especially in high-noise conditions and dialects with many unique vocabularies. By improving these aspects, the developed ASR model is expected to be more effective in recognizing the Mandar language across various dialects and recording conditions, making it widely applicable in local language technology applications.

**Comparison with Other ASR Models**

To provide a broader context for the model's performance, the fine-tuned Whisper was compared with two other ASR models, namely DeepSpeech and wav2vec 2.0, using the same test dataset.

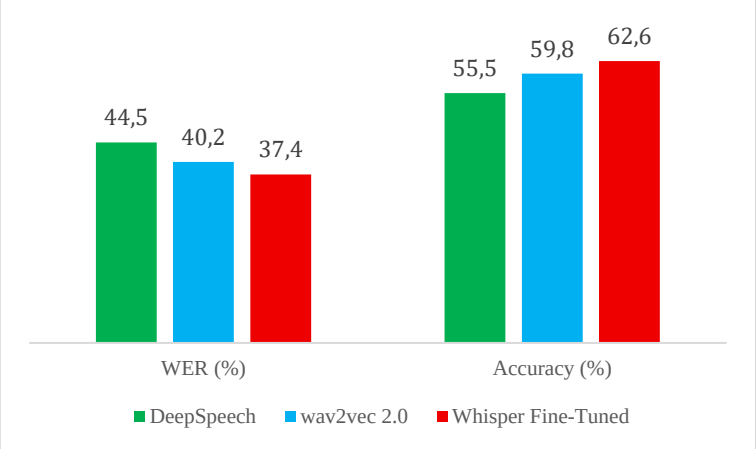


Fig. 3. WER and Accuracy Comparison Across ASR Models

Table 3 – Performance Comparison with Other ASR Models

| Model ASR | WER (%) | Accuracy (%) | Average Latency (seconds) |
|---|---|---|---|
| DeepSpeech | 44,5 | 55,5 | 2,8 |
| wav2vec 2.0 | 40,2 | 59,8 | 2,1 |
| Whisper Fine-Tuned | 37,4 | 62,6 | 1,5 |

**Comparative Analysis:**

- The fine-tuned Whisper model achieved the lowest WER compared to DeepSpeech and wav2vec 2.0, demonstrating superior performance in recognizing the Mandar language.

- In terms of latency, Whisper was faster, with an average processing time of 1.5 seconds per sentence, compared to the longer processing times required by DeepSpeech and wav2vec 2.0.
- This improved performance reflects the efficiency of Whisper's architecture and the benefits of fine-tuning with a localized dataset.

**Discussion and Research Implications**

The study's findings reveal that the fine-tuned Whisper model demonstrates significant advantages over both the zero-shot model and other ASR systems such as DeepSpeech and Wav2Vec 2.0. Fine-tuning enables the model to adapt more effectively to the specific linguistic characteristics of the Mandar language, including its distinctive phonetic structures, morphological forms, and vocabulary. These advantages indicate that Whisper substantially reduces the Word Error Rate (WER) and enhances transcription accuracy compared to models that are not retrained with language-specific data.

The superior performance of Whisper can be attributed to its multilingual pretraining and fine-tuning adaptability. Trained on over 680,000 hours of multilingual and multitask data, Whisper benefits from extensive cross-lingual phonetic and semantic representations. This multilingual foundation allows the model to generalize efficiently to low-resource languages like Mandar, whose phonetic features partially overlap with better-resourced languages in its pretraining corpus. During fine-tuning, the model recalibrates its attention mechanisms toward Mandar-specific speech patterns, achieving an optimal balance between transfer learning and localized adaptation. In contrast, DeepSpeech relies primarily on supervised learning with limited linguistic transfer capability, and Wav2Vec 2.0 focuses more on acoustic-level representation without multilingual contextual modeling.

In practical terms, Whisper's sequence-to-sequence Transformer architecture unifies acoustic and linguistic modeling within an end-to-end framework, reducing the error propagation typically found in modular ASR pipelines. This integration allows Whisper to outperform alternative architectures in both transcription accuracy and inference latency, making it highly suitable for real-time or interactive applications. The empirical results confirm this: Whisper achieved the lowest WER (37.4%) and highest accuracy (62.6%), outperforming Wav2Vec 2.0 (WER 40.2%) and DeepSpeech (WER 44.5%). These results validate the model's efficiency in adapting to underrepresented languages and diverse acoustic environments.

While fine-tuning produced substantial performance improvements, challenges remain in addressing dialectal variation and recording quality. Dialectal differences significantly affected transcription accuracy, with Standard Mandar (Dialect X) yielding the best results and lowest WER, while Bugis-influenced Mandar (Dialect Y) resulted in higher WER. This outcome highlights the model's limited exposure to underrepresented dialects, indicating a need for balanced data collection and dialect-specific adaptation strategies. Similarly, recording quality was found to be a critical factor in model performance. High-quality audio produced lower WER and higher accuracy, whereas noisy recordings increased WER and reduced accuracy. These findings underscore the importance of incorporating data augmentation and noise-robust training to improve generalization under varied acoustic conditions.

Despite these challenges, the study makes a valuable contribution to the development of ASR systems for low-resource languages. The results demonstrate that Whisper's multilingual pretraining, combined with efficient fine-tuning, aligns with the principles of Multilingual Transfer Learning (MTL) and Parameter-Efficient Transfer Learning (PETL). This hybrid strategy enables the model to achieve competitive accuracy even with a relatively small training corpus (≈36 hours), a crucial advantage for under-documented languages. Furthermore, Whisper's adaptability suggests strong potential for zero-shot and few-shot learning, which could further reduce reliance on large-scale annotated datasets in future ASR research.

The fine-tuned Whisper model's robust performance across metrics such as WER, accuracy, and latency positions it as a highly effective solution for speech recognition applications in local languages. Its implementation can be extended to diverse fields, including education, language preservation, and voice-based interaction systems for regional communication technologies.

**Limitations of the Study**

Despite the promising outcomes, several limitations must be acknowledged. First, the dataset size used for fine-tuning—approximately 36 hours of recorded speech—is relatively small compared to large-scale ASR benchmarks. This limited data volume may constrain the model's ability to generalize across unseen speakers, contexts, or rare phonetic patterns. Second, the imbalance of dialectal representation within the dataset resulted in varying performance, with the model achieving higher accuracy on Standard Mandar but lower accuracy on Bugis-influenced Mandar. This underperformance highlights the need for a more balanced and inclusive corpus to represent dialectal diversity. Third, although fine-tuning improved accuracy, there remains a risk of overfitting due to repeated exposure to specific recording environments and speaker profiles, potentially reducing robustness when applied to new data. Future iterations of the study should mitigate these limitations through larger-scale data collection, more aggressive regularization, and cross-dialectal validation.

**Potential of Zero-Shot and Few-Shot Learning**

To further address data scarcity challenges, zero-shot and few-shot learning approaches offer promising directions for future work. Whisper already exhibits zero-shot capabilities derived from its multilingual pretraining, enabling it to recognize untrained languages with minimal data. Building on this foundation, few-shot fine-tuning—where only a small number of labeled examples are used—could further reduce dependence on extensive data collection. Recent advances in parameter-efficient adaptation methods, such as LoRA (Low-Rank Adaptation) and adapter layers, allow large ASR models to adapt to new languages using limited computational resources and minimal labeled data. Implementing these techniques for Mandar and other regional languages would enhance model scalability, support faster deployment, and promote accessibility for other low-resource linguistic communities.

For future research, integrating advanced data augmentation techniques—including pitch variation, tempo modulation, and simulated environmental noise—can enhance resilience to recording variability. In addition, incorporating language model rescoring may further improve contextual understanding during decoding. Exploring parameter-efficient fine-tuning methods and few-shot adaptation can help reduce computational costs while maintaining high accuracy. Extending this research to other Austronesian languages could also generalize the findings and accelerate progress in low-resource speech recognition and language preservation efforts.

## 5. Conclusion

This study successfully enhanced speech recognition performance for the Mandar language through fine-tuning of the Whisper model using a locally collected dataset. The results show a significant improvement, reducing the *Word Error Rate* (WER) from 73.7% to 37.4% and increasing accuracy from 26.3% to 62.6%. The fine-tuned Whisper model also outperformed DeepSpeech and Wav2Vec 2.0, achieving the lowest WER and fastest latency. The optimized model demonstrates strong scalability and deployment potential across various domains, including education, local media transcription, and government documentation, indicating its readiness for real-world implementation. Collaboration with local linguistic communities and integration with national digital archives are recommended to enrich datasets and sustain long-term development. Beyond its technical contribution, this study holds policy implications for language preservation and regional AI development, supporting inclusive digital transformation and linguistic diversity in Indonesia. Future work should focus on dataset expansion, data augmentation, and few-shot learning to further enhance performance and extend the benefits of ASR technology for regional language preservation.

## References

Adnew, S., & Liang, P. P. (2024). *Semantically Corrected Amharic Automatic Speech Recognition*. https://arxiv.org/pdf/2404.13362

Ali, A., & Renals, S. (2018). Word error rate estimation for speech recognition: E-wer. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of*

*the Conference (Long Papers)*, *2*(2014), 20–24. https://doi.org/10.18653/v1/p18-2004

Arisaputra, P., Handoyo, A. T., & Zahra, A. (2024). XLS-R Deep Learning Model for Multilingual ASR on Low- Resource Languages: Indonesian, Javanese, and Sundanese. *ICIC Express Letters, Part B: Applications*, *15*(6), 551–559. https://doi.org/10.24507/icicelb.15.06.551

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.

Bawitlung, A., Dash, S. K., & Pattanayak, R. M. (2025). Mizo Automatic Speech Recognition: Leveraging Wav2vec 2.0 and XLS-R for Enhanced Accuracy in Low-Resource Language Processing. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *24*(7). https://doi.org/10.1145/3746063

Bender, E. M., Gebru, T., Mcmillan-Major, A., Shmitchell, S., & Shmitchell, S.-G. (2021). On the dangers of stochastic parrots: Can language models be too big. *Dl.Acm.Org*, 610–623. https://doi.org/10.1145/3442188.3445922

Bhat, C., & Strik, H. (2025). Two-stage data augmentation for improved ASR performance for dysarthric speech. *Computers in Biology and Medicine*, *189*, 109954. https://doi.org/10.1016/J.COMPBIOMED.2025.109954

Brodkin, D. (2022). Two steps to high absolutive syntax: Austronesian voice and agent focus in Mandar. *Journal of East Asian Linguistics*, *31*(4), 465–516. https://doi.org/10.1007/S10831-022-09248-0/METRICS

Cahyawijaya, S., Lovenia, H., Aji, A. F., Winata, G. I., Wilie, B., Koto, F., Mahendra, R., Wibisono, C., Romadhony, A., Vincentio, K., Santoso, J., Moeljadi, D., Wirawan, C., Hudi, F., Wicaksono, M. S., Parmonangan, I. H., Alfina, I., Putra, I. F., Rahmadani, S., … Purwarianti, A. (2023). NusaCrowd: Open Source Initiative for Indonesian NLP Resources. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 13745–13818. https://doi.org/10.18653/V1/2023.FINDINGS-ACL.868

Das, R., & Singh, T. D. (2023). Multimodal Sentiment Analysis: A Survey of Methods, Trends, and Challenges. *ACM Computing Surveys*, *55*(13). https://doi.org/10.1145/3586075

DeySpandan, SahidullahMd, & SahaGoutam. (2022). An Overview of Indian Spoken Language Recognition from Machine Learning Perspective. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *21*(6), 1–45. https://doi.org/10.1145/3523179

Ferdiansyah, D., & Aditya, C. S. K. (2024). Implementasi Automatic Speech Recognition Bacaan Al-Qur'an Menggunakan Metode Wav2Vec 2.0 dan OpenAI-Whisper. *Jurnal Teknik Elektro Dan Komputer TRIAC*, *11*(1), 11–16. https://doi.org/10.21107/triac.v11i1.24332

Gillioz, A., Casas, J., … E. M.-2020 15th C., & 2020, U. (2020). Overview of the Transformer-based Models for NLP Tasks. *Ieeexplore.Ieee.Org*. https://doi.org/10.15439/2020F20

Gurunath Shivakumar, P., & Georgiou, P. (2020). Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. *Computer Speech & Language*, *63*, 101077. https://doi.org/10.1016/J.CSL.2020.101077

Hasrullah. (2018). *Ribuan Mandar: Preservasi Bahasa Dan Kebudayaan Mandar Sulawesi Barat (Rancangan Aplikasi Berbasis Android Dengan Sistem Sociopreneurship)*. https://www.researchgate.net/publication/331630308_Ribuan_Mandar_Preservasi_Bahasa_Dan_Kebudayaan_Mandar_Sulawesi_Barat_Rancangan_Aplikasi_Berbasis_Android_Dengan_Sistem_Sociopreneurship

Imam, S. H., Belay, T. D., Husse, K. Y., Ahmad, I. S., Abdulmumin, I., Umar, H. A., Bello, M. Y., Nakatumba-Nabende, J., Yimam, S. M., & Muhammad, S. H. (2025). *Automatic Speech Recognition (ASR) for African Low-Resource Languages: A Systematic Literature Review*. *1*. https://arxiv.org/pdf/2510.01145

Kastner, K., Wang, G., Elias, I., Saeki, T., Mengibar, P. M., Beaufays, F., Rosenberg, A., & Ramabhadran, B. (2025). Speech Re-Painting for Robust ASR. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. https://doi.org/10.1109/ICASSP49660.2025.10888357

Kuhn, K., Kersken, V., Reuter, B., Egger, N., & Zimmermann, G. (2023). Measuring the accuracy of automatic speech recognition solutions. *Dl.Acm.Org*, *16*(4).

https://doi.org/10.1145/3636513

Lei, C., Singh, S., Hou, F., & Wang, R. (2024). Mix-fine-tune: An Alternate Fine-tuning Strategy for Domain Adaptation and Generalization of Low-resource ASR. *Proceedings of the 6th ACM International Conference on Multimedia in Asia, MMAsia 2024*. https://doi.org/10.1145/3696409.3700259

Liu, Y., Yang, X., & Qu, D. (2024a). Exploration of Whisper fine-tuning strategies for low-resource ASR. *Eurasip Journal on Audio, Speech, and Music Processing*, *2024*(1), 1–11. https://doi.org/10.1186/S13636-024-00349-3/FIGURES/6

Liu, Y., Yang, X., & Qu, D. (2024b). Exploration of Whisper fine-tuning strategies for low-resource ASR. *Eurasip Journal on Audio, Speech, and Music Processing*, *2024*(1), 1–11. https://doi.org/10.1186/S13636-024-00349-3/FIGURES/6

Machado, F., Rahali, A., & Akhloufi, M. A. (2023). End-to-end transformer-based models in textual-based NLP. *Mdpi.Com*. https://doi.org/10.3390/ai4010004

Nakatumba-NabendeJoyce, KagumireSulaiman, KantonoCaroline, & NabendePeter. (2025). A Systematic Literature Review on Bias Evaluation and Mitigation in Automatic Speech Recognition Models for Low-Resource African Languages. *ACM Computing Surveys*. https://doi.org/10.1145/3769089

Nalli, S., Haria, S., Hill, M. D., Swift, M. M., Volos, H., & Keeton, K. (2017). An Analysis of Persistent Memory Use with WHISPER. *ACM SIGPLAN Notices*, *52*(4), 135–148. https://doi.org/10.1145/3093336.3037730

Niu, T., Chen, Y., Qu, D., & Hu, H. (2025). Enhancing Far-Field Speech Recognition with Mixer: A Novel Data Augmentation Approach. *Applied Sciences 2025, Vol. 15, Page 4073*, *15*(7), 4073. https://doi.org/10.3390/APP15074073

Nugroho, K., Noersasongko, E., Purwanto, Muljono, & Setiadi, D. R. I. M. (2022). Enhanced Indonesian Ethnic Speaker Recognition using Data Augmentation Deep Neural Network. *Journal of King Saud University - Computer and Information Sciences*, *34*(7), 4375–4384. https://doi.org/10.1016/J.JKSUCI.2021.04.002

Nurfadhilah, E., Yuyun, Santosa, A., Latief, A. D., Nurul Afra, D. I., Gusnawaty, Pammuda, Kaharuddin, M. N., Rosvita, I., Nurfaedah, & Hazriani. (2024). Comparative Analysis of Part of Speech Tagging Methods for the Bugis Language: From Statistical to Deep Neural Approaches. *International Conference on Computer, Control, Informatics and Its Applications*, *IC3INA*, *2024*, 48–53. https://doi.org/10.1109/IC3INA64086.2024.10732773

Papala, G., Ransing, A., and, P. J.-S. C. P., & 2023, undefined. (2023). Sentiment Analysis and Speaker Diarization in Hindi and Marathi Using using Finetuned Whisper: Sentiment Analysis in Hindi and Marathi. *Scpe.Org*, *24*(4), 835–846. https://doi.org/10.12694/scpe.v24i4.2248

Peng, J., Stafylakis, T., Gu, R., Plchot, O., Mošner, L., Burget, L., & Černocký, J. (2023). Parameter-Efficient Transfer Learning of Pre-Trained Transformer Models for Speaker Verification Using Adapters. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2023-June*. https://doi.org/10.1109/ICASSP49357.2023.10094795

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings.Mlr.Press*. https://proceedings.mlr.press/v202/radford23a.html

Rolland, T., Abad, A., Cucchiarini, C., & Strik, H. (2022). *Multilingual Transfer Learning for Children Automatic Speech Recognition* (pp. 7314–7320). https://aclanthology.org/2022.lrec-1.795/

Romero, M., Gómez-Canaval, S., & Torre, I. G. (2024). Automatic Speech Recognition Advancements for Indigenous Languages of the Americas. *Applied Sciences (Switzerland)*, *14*(15), 6497. https://doi.org/10.3390/APP14156497/S1

Roos, Q. (2022). *Fine-tuning pre-trained language models for CEFR-level and keyword conditioned text generation: Acomparison between Google's T5 and OpenAI's GPT-2*. https://www.diva-portal.org/smash/record.jsf?pid=diva2:1708538

Rusdiah, R., Rasjid, N., Irianti, A., Adiheri, A., & Reski, R. (2023). Digitalisasi Cerita Rakyat

Mandar. *Jurnal Pengabdian Masyarakat Universitas Lamappapoleonro*, *1*(2), 66–70. https://jurnal.abdimas.unipol.ac.id/index.php/pengabdian-jurnal/article/view/17

Sekiguchi, K., Bando, Y., Audio, A. N.-… on, Speech, U., & 2019, U. (2019). Semi-supervised multichannel speech enhancement with a deep speech prior. *Ieeexplore.Ieee.Org*. https://ieeexplore.ieee.org/abstract/document/8861142/

Serva, M., Pasquini -, M., Tawaqal, B., & Suyanto, S. (2021). Recognizing Five Major Dialects in Indonesia Based on MFCC and DRNN. *Journal of Physics: Conference Series*, *1844*(1), 012003. https://doi.org/10.1088/1742-6596/1844/1/012003

Suyanto, S., Arifianto, A., Sirwan, A., & Rizaendra, A. P. (2020). End-to-End Speech Recognition Models for a Low-Resourced Indonesian Language. *2020 8th International Conference on Information and Communication Technology, ICoICT 2020*. https://doi.org/10.1109/ICOICT49345.2020.9166346

Xiao, A., Zheng, W., Keren, G., Le, D., Zhang, F., Fuegen, C., Kalinli, O., Saraf, Y., & Mohamed, A. (2021). Scaling ASR Improves Zero and Few Shot Learning. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2022-September*, 5135–5139. https://doi.org/10.21437/Interspeech.2022-11023

Zevallos, R., Cordova, J., & Camacho, L. (2020). Automatic Speech Recognition of Quechua Language Using HMM Toolkit. *Communications in Computer and Information Science*, *1070 CCIS*, 61–68. https://doi.org/10.1007/978-3-030-46140-9_6

Zhao, J., & Zhang, W. Q. (2022). Improving Automatic Speech Recognition Performance for Low-Resource Languages with Self-Supervised Models. *IEEE Journal on Selected Topics in Signal Processing*, *16*(6), 1227–1241. https://doi.org/10.1109/JSTSP.2022.3184480