Fine-tuning Whisper Tiny for Swahili ASR: Challenges and Recommendations for Low-Resource Speech Recognition

Avinash Kumar Sharma, Manas R Pandya, Arpit Shukla
{zda23m011, zda23b019, zda23m007}@iitmz.ac.in
Indian Institute of Technology Madras, Zanzibar Campus

Abstract

Automatic Speech Recognition (ASR) technologies have seen significant advancements, yet many widely spoken languages remain underrepresented. This paper explores the fine-tuning of OpenAI's Whisper Tiny model (39M parameters) for Swahili, a lingua franca for over 100 million people across East Africa. Using a dataset of 5,520 Swahili audio samples, we analyze the model's performance, error patterns, and limitations after fine-tuning. Our results demonstrate the potential of fine-tuning for improving transcription accuracy, while also highlighting persistent challenges such as phonetic misinterpretations, named entity recognition failures, and difficulties with morphologically complex words. We provide recommendations for improving Swahili ASR, including scaling to larger model variants, architectural adaptations for agglutinative languages, and data enhancement strategies. This work contributes to the growing body of research on adapting pre-trained multilingual ASR systems to low-resource languages, emphasizing the need for approaches that account for the unique linguistic features of Bantu languages.

1 Introduction

Automatic Speech Recognition (ASR) has revolutionized human-computer interaction, but many widely spoken languages, including Swahili, remain underrepresented in ASR technology. Swahili (Kiswahili) is a lingua franca for over 100 million people across East Africa, yet its presence in modern speech recognition systems is minimal compared to high-resource languages like English, Mandarin, and Spanish.

This study explores the fine-tuning of OpenAI's Whisper Tiny model - a lightweight variant of the Whisper ASR system - for Swahili speech recognition. Whisper, trained on an extensive 680k hour multilingual dataset, shows promise across multiple languages but requires targeted adaptation for low-resource languages like Swahili to perform effectively. We investigate the impact of fine-tuning on model performance, focusing on transcription accuracy, error patterns, and the challenges of adapting compact ASR models for low-resource languages.

To guide this research, we aim to answer the following key questions:
How does fine-tuning the Whisper Tiny model (39M parameters) improve its performance for Swahili speech recognition? What are the main transcription errors observed in the fine-tuned model, and how can they be addressed? What are the challenges of using compact ASR models for low-resource languages, and how can they be mitigated through training strategies and architectural adaptations?

By addressing these questions, this work contributes to the ongoing effort to improve ASR for languages with limited digital resources, paving the way for more inclusive speech technologies. Our findings highlight the importance of fine-tuning for low-resource languages and provide actionable recommendations for future research and development in Swahili ASR.

2 Background and Related Work

2.1 Automatic Speech Recognition for Low-Resource Languages

Developing ASR systems for low-resource languages like Swahili comes with significant challenges, primarily due to the scarcity of transcribed speech data. Most state-of-the-art ASR models rely on large-scale paired audio-text datasets, which are often unavailable for such languages. Recent advances in transfer learning have enabled the adaptation of models pre-trained on high-resource languages, allowing ASR systems to perform well even with limited native-language data (Besacier et al., 2014). Fine-tuning these models is critical

Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 74–81
July 31, 2025 ©2025 Association for Computational Linguistics

to achieving optimal performance, as it allows the system to adapt to the unique phonetic and linguistic characteristics of the target language. Our study explores this process by fine-tuning the Whisper Tiny model for Swahili, evaluating its performance and identifying areas for improvement.

2.2 The Whisper ASR System

OpenAI's Whisper model represents a major step forward in multilingual ASR Radford et al. (2023). Trained on a diverse dataset of 680,000 hours of labeled audio, it offers robust transcription capabilities across multiple languages. The Whisper family consists of models of varying sizes, ranging from Tiny (39M parameters) to Large (1.5B parameters), each balancing computational efficiency and transcription accuracy. Researchers have successfully fine-tuned Whisper for several low-resource languages, such as Amharic (Abdou Mohamed et al., 2024), Yoruba (Ahia et al., 2024), and Nepali (Ghimire et al., 2024). Despite these advancements, comprehensive studies on Whisper's adaptation for Swahili remain limited. Our work addresses this gap by evaluating the impact of fine-tuning on Swahili ASR performance, providing insights into the challenges and opportunities of adapting compact ASR models for low-resource languages.

While ASR for major languages has seen substantial advances, Swahili ASR development remains limited despite its widespread use. Recent efforts include work by Tunde-Onadele and Chao (2022) who developed initial speech recognition models for Swahili using traditional Hidden Markov Model approaches. In the neural era, several pre-trained models have been adapted for Swahili, including wav2vec 2.0 variants (Akash, 2023; SpeechBrain Team, 2022) which leverage self-supervised learning on unlabeled audio. These efforts, however, often lack detailed documentation of the fine-tuning process and error analysis specific to Swahili's linguistic characteristics. Our work complements these efforts by providing a systematic analysis of fine-tuning a compact transformer-based model and documenting language-specific challenges.

3 Methodology

3.1 Dataset Characteristics

This study utilized a comprehensive Swahili speech dataset comprising 5,520 audio samples with corresponding transcriptions.

The dataset was sourced from KenSpeech (Awino et al., 2022), and from an initial pool of approximately 6,300 samples. The audio files were scraped in batches with appropriate processing delays to ensure data integrity. Each audio file was paired with its corresponding transcript, creating a parallel corpus suitable for ASR model training.

We filtered transcripts using automatic language detection via the `langdetect` library. Transcripts predominantly identified as Swahili were retained, while those detected as primarily English were excluded. This filtering process ensured the linguistic purity necessary for training a robust Swahili ASR system.

3.1.1 Length Distributions

The dataset's transcript lengths follow a roughly normal distribution, with a slight negative skew. The majority of samples (69.7%) contain between 21-40 words, indicating a prevalence of medium-length utterances suitable for ASR training. Figure 1 shows the distribution of transcript lengths by category.

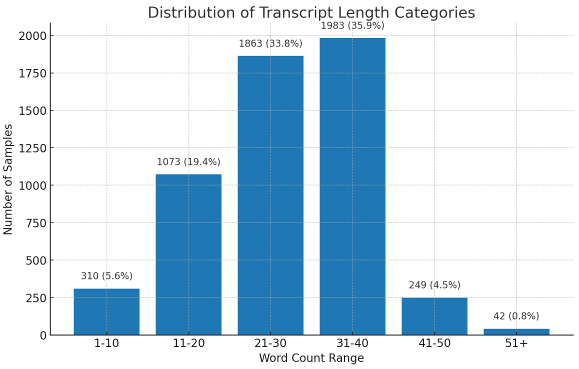

Figure 1: Distribution of Transcript Length Categories showing the number of samples in each word count range. The 21-30 and 31-40 word ranges contain the majority of samples (69.7%).

The detailed frequency distribution of word counts is illustrated in Figure 2, showing the mean and median values.

The tokenized transcripts exhibit a similar distribution pattern but with higher values, as illustrated in Figure 3. This reflects the nature of subword tokenization, where individual words are often split into multiple tokens, particularly in morphologically rich languages like Swahili.

Through the statistical analysis, we find that the Swahili transcriptions averaged 27.23 words and 71.5 tokens, with maximums of 67 words and 170 tokens; medians were slightly higher than means.

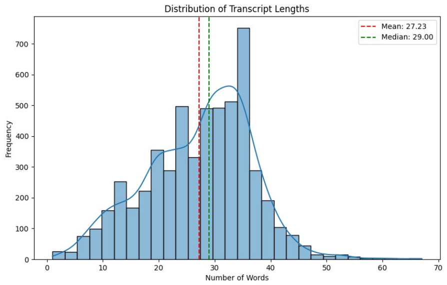

Figure 2: Histogram of Transcript Lengths showing the frequency distribution of word counts. The mean (27.23 words) and median (29.00 words) are indicated by vertical dashed lines.

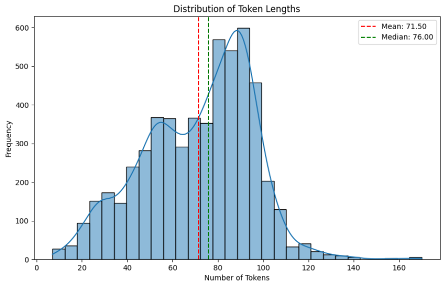

Figure 3: Distribution of Token Lengths across the dataset. The mean (71.50 tokens) and median (76.00 tokens) are indicated by vertical dashed lines, showing the increased granularity when words are tokenized.

3.1.2 Data Partitioning

To facilitate model training and evaluation, the dataset was partitioned using a standard 80/20 split, resulting in a Training set of 4,416 samples (80%), and a Validation set of 1,104 samples (20%).

This partitioning strategy ensures sufficient data for model training while retaining an adequate portion for validation to assess generalization performance. The stratified splitting approach maintained similar transcript length distributions across both sets to prevent evaluation bias.

3.2 Model Architecture[1]

This study employed OpenAI's Whisper Tiny model as the foundation for Swahili ASR development. The Whisper family of models represents a state-of-the-art approach to multilingual speech recognition, with variants ranging from Tiny (39M parameters) to Large (1.5B parameters).

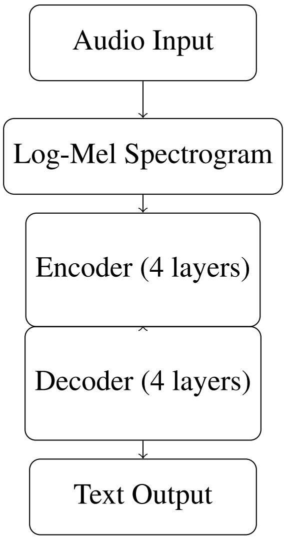

Figure 4: Simplified diagram of the Whisper model architecture, showing the processing pipeline from audio input to text output through the encoder-decoder transformer structure.

3.2.1 Whisper Tiny Specifications

The Whisper Tiny model (OpenAI, 2023), selected for its computational efficiency while maintaining reasonable performance, features the architectural specifications shown in Table 1.

Component	**Specification**
Model type	Encoder-decoder transformer
Parameter count	39 million
Encoder layers	4
Decoder layers	4
Model dimension	384
Attention heads	6
Audio feature extractor	CNN
Maximum context length	3,000 frames (≈ 30s)

Table 1: Architectural Specifications of Whisper Tiny Model

Compared to larger Whisper variants, the Tiny model offers substantially reduced computational requirements while retaining the core architectural elements that enable effective speech recognition. Figure 4 illustrates the overall architectural design of the Whisper model.

3.2.2 Audio Processing Pipeline

The Whisper architecture processes audio through a multi-stage pipeline designed to efficiently

[1] The configuration is in the Appendix A

transform raw audio into transcribed text:

Audio Preprocessing: Raw audio is resampled to 16 kHz and converted into 80-channel log-Mel spectrograms with 25ms windows and 10ms stride.
Encoder Processing: The encoder, consisting of 4 transformer layers with 4 attention heads each, processes these spectrograms to create contextual audio representations. This stage captures the acoustic and phonetic features of the input speech.
Decoder Generation: The decoder, also comprising 4 transformer layers, generates text tokens autoregressively based on the encoded representations. The decoder incorporates cross-attention mechanisms that attend to the encoder outputs, enabling the model to align speech features with textual elements.
Token Prediction: At each decoding step, the model predicts the next token from a multilingual vocabulary of approximately 50,000 tokens, which includes subword units for Swahili and other languages.

3.2.3 Adaptation for Swahili

While the base Whisper Tiny model includes some support for Swahili through its multilingual pretraining, specific adaptations were implemented to enhance its performance:

Language-Specific Initialization: The decoder was initialized with Swahili language ID tokens to bias generation toward Swahili output.

Task Configuration: The model was configured specifically for transcription tasks rather than translation, focusing its capabilities on accurate within-language processing.

4 Results

4.1 Training Progress

The model was trained for both 5 and 100 epochs, with evaluation metrics captured at intermediate and final stages. The 5-epoch training proceeded without significant technical issues, though some audio processing challenges were observed. The extended 100-epoch training demonstrated substantial improvements, with validation loss decreasing steadily and WER improving from 82.95% to 30.62%, highlighting the benefits of extended fine-tuning for low-resource languages.

Compared to the small Whisper model trained by Pplantiga on Hugging Face (Plantinga, 2025), which achieved a WER of 27.62%, our tiny model, with a WER of 30.62%, offers a better trade-off by significantly reducing complexity, even though the performance is slightly lower.

4.2 Training and Validation Dynamics

Table 2 summarizes the Training and Validation metrics for both 5-epoch and 100-epoch training regimes.

Metric	5 Epochs	100 Epochs
Training Loss	0.8426	0.0001
Validation Loss	1.2317	1.5977
WER	82.95%	30.62%

Table 2: Metrics for 5-Epoch and 100-Epoch Training

The 100-epoch model achieved a WER of 30.62%, representing a 63.1% improvement over the 5-epoch model. This reduction in WER underscores the importance of extended training for low-resource languages, though challenges remain in achieving practical usability (typically <20% WER).

4.3 Qualitative Error Analysis

A detailed examination of model outputs revealed several recurring error patterns, with notable improvements in the 100-epoch model.

4.3.1 Phonetic Approximations

The model frequently produced phonetically plausible but incorrect transcriptions, as shown in Table 3. While the 100-epoch model reduced errors like double vowel omissions and segmentation issues, challenges persisted.

Reference	5-Epoch Prediction	100-Epoch Prediction
muunganisho wa maarifa asilia na sayansi ya kisasa utaleta jitihada ya ubindamu kuleta amani na mazingira yetu ushauri wangu	m**u**nganisho wa maarifa asilia na sayansi **za** **g**isasa utali**t**a jiti**a**da ya ubi**na** **damu** kuleta amani na mazi**ngi** **ra**hetu ushauri wangu	muunganisho wa maarifa asilia na sayansi ya kisasa utaleta jitihada ya ubindamu kuleta amani na mazingira yetu ushauri wangu

Table 3: Example of Phonetic Approximation Errors

4.3.2 Named Entity Recognition

The model exhibited difficulty with proper nouns, as demonstrated in Table 4. The 100-epoch model showed slight improvements but still struggled with non-Swahili names.

Reference	5-Epoch Prediction	100-Epoch Prediction
moto shuleni bweni la shule ya st brigid’s kiminini liliteketea jana usiku hakuna mwanafunzi aliyejeruhiwa lakini uchunguzi umeaanza	moto **shule ya ni** bweni la shule ya **sandbridge it kimi mini** ilili**t**aketea jana usiku hakuna mwanafunzi aliye jeruhiwa **na kini** uchunguzi ume**han za**	moto shuleni bweni la shule ya st brigid’s kiminini liliteketea jana usiku hakuna mwanafunzi aliyejeruhiwa lakini uchunguzi umeaanza

Table 4: Example of Named Entity Recognition Errors

4.3.3 Repetitive Pattern Generation

In several instances, the model produced repetitive, non-informative output, particularly when encountering challenging audio, as shown in Table 5. The 100-epoch model reduced the frequency of such errors but did not eliminate them entirely.

Reference	5-Epoch Prediction	100-Epoch Prediction
murang’a viongozi wa kidini wazidiwa na hisia hasa wanapohudhuria maombolezo	**mwanamke bomba tunawakata wakati wa kisii kwa kama kama kama kama kama kama kama kama kama...**	murang’a viongozi wa kidini wazidiwa na hisia **hasira waziri waadi wa wadhaingizaa kutangaza**

Table 5: Example of Repetitive Pattern Generation

4.3.4 Sentence Length Correlation

Analysis indicated that transcription accuracy had a moderate negative correlation with sentence length, with longer sentences generally exhibiting higher error rates. This trend persisted in the 100-epoch model, though the magnitude of the correlation was slightly reduced.

5 Discussion

5.1 Model Capacity Limitations

The high Word Error Rate (WER) observed in the 5-epoch model (82.95%) suggests that the Whisper Tiny model’s capacity (39M parameters) is insufficient for robust Swahili ASR. The 100-epoch model achieved a WER of 30.62%, demonstrating significant improvement but still falling short of practical usability.

5.2 Linguistic Challenges in Swahili ASR

Several linguistic features of Swahili present particular challenges for ASR systems:

5.2.1 Morphological Complexity

As an agglutinative language, Swahili constructs words by combining multiple morphemes. A single Swahili verb can encode information about subject, object, tense, aspect, and mood through prefixes and suffixes. This complexity increases the vocabulary space and makes word boundary detection challenging, as evidenced by errors like splitting *jitihada* into *jiti ada*. The 100-epoch model showed slight improvements in handling morphologically complex words but still struggled with segmentation.

5.2.2 Tonal Variations

While Standard Swahili is not strictly tonal, variations in intonation can affect meaning, particularly in regional dialects. The model’s difficulty with certain phonetic distinctions may be partially attributed to inadequate representation of tonal features. Extended training reduced some phonetic errors but did not fully address this challenge.

5.2.3 Dialectal Diversity

Swahili exhibits significant dialectal variation across East Africa, with notable differences between Tanzanian and Kenyan varieties. The dataset’s regional representation may impact the model’s ability to generalize across dialectal boundaries. The 100-epoch model showed improved generalization but still struggled with dialect-specific variations.

5.3 Data Considerations

The size of the training dataset (5,520 samples ≈ 19.7 hours) is relatively small compared to datasets used for high-resource language ASR development, which often include tens of thousands of hours of audio. Table 6 provides a comparison with other ASR datasets.

Dataset	Size	Language
Current Study	19.7 hours	Swahili
LibriSpeech	960 hours	English
Common Voice	1,400+ hours	Multiple lang.
High-resource ASR	10,000+ hours	Various

Table 6: Comparison of Dataset Sizes for ASR Development

The limited data diversity may constrain the model's ability to generalize to varied speakers, acoustic environments, and linguistic contexts. Extended training improved performance but did not fully compensate for the dataset's limitations.

6 Future Work

Based on our findings from both 5-epoch and 100-epoch fine-tuning, we propose several recommendations for improving Swahili ASR performance:

6.1 Model Architecture Enhancements

Table 7 summarizes our recommendations for scaling to larger model variants, which could further reduce the WER observed in our experiments.

1. Scale to Larger Model Variants	
The most immediate improvement would likely come from utilizing larger Whisper models:	
Base	74M parameters - Recommended minimum for Swahili ASR, with expected WER improvements of 30-50% over Tiny.
Small	244M parameters - Optimal balance of performance and efficiency, often achieving WERs below 30%.
Medium	769M parameters - For research contexts with sufficient computational resources, offering potential for further improvement.

Table 7: Model Scaling Recommendations

Table 8 presents our recommendations for architectural adaptations to better handle Swahili's linguistic features.

2. Architectural Adaptations	
Consider modifications to the base Whisper architecture to better accommodate Swahili's linguistic features:	
Tokenization	Enhanced subword tokenization specifically designed for agglutinative languages to address segmentation errors.
Attention	Augmented attention mechanisms to better capture long-range dependencies in morphologically complex words.
Audio Features	Additional acoustic feature extraction layers to better represent tonal variations and phonetic nuances.

Table 8: Architectural Adaptation Recommendations

6.2 Training Methodology Improvements

1. **Extended training duration**: Our 100-epoch training demonstrated significant WER improvements, suggesting that longer training durations are beneficial for low-resource languages. Early stopping based on validation performance can prevent overfitting.

2. **Learning rate scheduling**: Implement more sophisticated learning rate schedules, such as cosine annealing with warm restarts, to improve optimization dynamics and convergence.

3. **Hyperparameter optimization**: Conduct systematic grid search or Bayesian optimization of key hyperparameters, including batch size and learning rate, to maximize performance.

4. **Progressive training**: Implement a curriculum learning approach where the model initially trains on shorter, simpler utterances before progressing to more complex examples, as longer sentences remain challenging.

6.3 Data Enhancement Strategies

Table 9 outlines our recommended data enhancement strategies to address the limitations of the current dataset.

7 Conclusion

This study evaluated fine-tuning the Whisper Tiny model for Swahili ASR with 5-epoch and 100-epoch training regimes. The 100-epoch model

Strategy	Approach
Data augmentation	Time stretching, pitch shifting, background noise, room impulse response, speed perturbation (0.9x, 1.0x, 1.1x)
Data quality	Review audio-transcript pairs for misalignments or errors.
Dialectal balancing	Represent major Swahili dialects to improve regional generalization.
Additional sources	Public archives, parliamentary proceedings, educational materials, user recordings.

Table 9: Recommended Data Enhancement Strategies

achieved a WER of 30.62%, significantly better than the 5-epoch model (82.95%). However, performance remains below practical utility, highlighting challenges in adapting compact ASR models for low-resource languages.

Error patterns, such as phonetic approximations and difficulties with named entity recognition and dialectal variations, suggest limitations in capturing Swahili's linguistic complexities. While extended training reduced some errors, issues with morphologically complex words persist.

This research contributes to adapting multilingual ASR systems for low-resource languages and emphasizes the need for approaches tailored to Bantu languages like Swahili. The work should focus on:
Evaluating larger Whisper models for scalability. Developing architectural adaptations for Swahili and Bantu languages. Creating more diverse Swahili speech datasets. Exploring multitask learning approaches. Developing ASR systems for languages like Swahili is a crucial step toward more inclusive speech technology for diverse linguistic communities.

Limitations

This study has several limitations. First, only the Whisper Tiny model was used, which may not represent the performance of larger Whisper variants for Swahili ASR. Second, the dataset, with 5,520 samples, is small, potentially limiting exposure to diverse Swahili speech patterns. Third, we focused on the standard Swahili dialect without accounting for regional variations. Fourth, further hyperparameter optimization could improve results, despite training for 100 epochs. Finally, our evaluation based on Word Error Rate may not fully reflect the semantic accuracy or practical usability of the transcriptions.

References

Noreen Abdou Mohamed, Arbi Allak, Karim Gaanoun, Ibrahim Benelallam, Zineb Erraji, and Ahmed Bahafid. 2024. Multilingual speech recognition initiative for african languages. *International Journal of Data Science and Analytics*, pages 1–16.

Orevaoghene Ahia, Ayeni Aremu, David Abagyan, Hila Gonen, David I. Adelani, David Abolade, Noah A. Smith, and Yulia Tsvetkov. 2024. Voices unheard: Nlp resources and models for yorúbá regional dialects. *arXiv preprint arXiv:2406.19564*.

P. B. Akash. 2023. Xlsr swahili model. https://huggingface.co/Akashpb13/Swahili_xlsr.

David Awino, Lawrence Muchemi, Lilian D. A. Wanzare, Evans Ombui, Bernard Wanjawa, Ochieng McOnyango, and Florence Indede. 2022. Kenspeech: Swahili speech transcriptions. https://doi.org/10.7910/DVN/YHXJSU.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Ram Rijal Ghimire, Prajwal Poudyal, and Bal Krishna Bal. 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)*, pages 408–415.

OpenAI. 2023. Whisper: A universal speech recognition model. [Accessed: 9-Mar-2025].

Peter Plantinga. 2025. Whisper small model for swahili.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

SpeechBrain Team. 2022. Asr wav2vec2 model for swahili. https://huggingface.co/speechbrain/asr-wav2vec2-dvoice-swahili.

Olamide Tunde-Onadele and Joseph Chao. 2022. Speech recognition for low-resource languages: Case study on swahili. In *Proceedings of the 3rd Workshop on Computational Methods for Endangered Languages*, pages 148–152.

Appendix

A Fine-Tuning Configuration

The adaptation of the Whisper Tiny model for Swahili ASR was accomplished through a targeted fine-tuning process optimized for efficient learning while preventing overfitting. Table 10 details the hyperparameters selected for this fine-tuning procedure.

Hyperparameter	Value
Batch size	8
Gradient accumulation steps	4
Effective batch size	32 (8 × 4)
Learning rate	5e-5
Learning rate scheduler	Linear with warmup
Warmup ratio	0.1
Number of epochs	5 / 100
Optimizer	AdamW
Weight decay	0.01
Mixed precision	FP16 (enabled)
Gradient checkpointing	Enabled
Language configuration	Forced decoder IDs set to Swahili
Maximum generation length	225 tokens

Table 10: Hyperparameters for Whisper Tiny Fine-tuning

The selection of these hyperparameters was guided by several considerations:

Memory Optimization: The combination of a moderate batch size (8) with gradient accumulation steps (4) yielded an effective batch size of 32, balancing between statistical efficiency and GPU memory constraints.

Learning Dynamics: The learning rate of 5e-5 with a warmup ratio of 0.1 was chosen to allow gradual adaptation of the pre-trained weights while avoiding destructive updates early in training.

Computational Efficiency: Mixed precision training (FP16) and gradient checkpointing were enabled to optimize GPU memory usage and accelerate training without sacrificing model quality.

Language Specificity: Forced decoder IDs ensured that the model generated Swahili text regardless of the detected language in the audio, which was critical for focused adaptation.

A.1 Training Infrastructure

The fine-tuning was conducted on a high-performance computing environment with the specifications shown in Table 11.

Component	Specification
GPU	NVIDIA A100
GPU Memory	16GB VRAM
Training Framework	PyTorch 1.13 HuggingFace Transformers
Training Duration	34 minutes (5 epochs) 11 hours (100 epochs)
Total Training Steps	690 (5 epochs) 13,500 (100 epochs)

Table 11: Training Infrastructure Specifications