Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Design and Development of Afan Oromo Speech Recognition and integrating with Dialogue System Using Whisper and RASA

Teklu Urgessa
Adama Science and Technology University
Worku Jifara
Adama Science and Technology University
Adugna Deksiso
adugna.deksiso@astu.edu.et

Adama Science and Technology University

Research Article

Keywords: Automatic Speech Recognition, Afan Oromo, Whisper, Low-Resource Languages, Fine-Tuning, RASA, SDS, Conversational AI

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10010573/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Automatic Speech Recognition (ASR) for low-resource languages remains a significant challenge due to limited transcribed data and the lack of systematic guidance on model selection. This paper presents the development of an Afan Oromo ASR system using fine-tuned Whisper-small, integrated with a RASA-based conversational agent. We curated a 76.88-hour corpus, implemented persistent dataset splits, applied data augmentation, and developed a robust checkpointing system. The fine-tuned model achieved a Word Error Rate (WER) of 16.88% on the validation set and 17.42% on the held-out test set. We further detail the integration of this model with RASA, creating a complete speech-to-intent pipeline. Our results demonstrate the viability of state-of-the-art ASR for Afan Oromo and provide a practical template for deploying conversational AI in low-resource settings.

1. Introduction

Automatic Speech Recognition (ASR) has become a fundamental component of modern human–computer interaction, enabling applications such as virtual assistants, automated transcription systems, accessibility tools, and conversational agents. Despite remarkable progress in ASR for high-resource languages such as English and Mandarin, developing robust speech technologies for low-resource languages remains a significant challenge due to limited speech corpora, linguistic resources, and computational infrastructure [1].

Afan Oromo (Oromo) is a Cushitic language belonging to the Afro-asiatic language family and is spoken by more than 40 million people in Ethiopia and neighboring countries [2]. Despite its large number of speakers, Afan Oromo remains underrepresented in speech and language technologies. Existing ASR research for the language is limited, and the performance of available systems remains substantially below that achieved for high-resource languages. One of the most notable previous studies developed an Afan Oromo Large Vocabulary Continuous Speech Recognition system based on a Speech Transformer architecture and reported a Word Error Rate (WER) of 40.2% on a broadcast news corpus [3].

Recent advances in deep learning and transfer learning have transformed the ASR landscape. In particular, OpenAI's Whisper model, trained on approximately 680,000 hours of multilingual and multitask speech data, has demonstrated strong cross-lingual speech recognition capabilities and robust performance in low-resource settings. Several studies have shown that fine-tuning pre-trained Whisper models on relatively small language-specific datasets can significantly improve recognition accuracy while reducing training costs [4], [5].

The growing availability of African speech datasets and multilingual speech models has further accelerated research on low-resource ASR. Benchmarking studies across multiple African languages have demonstrated the effectiveness of transformer-based architectures such as Whisper, XLS-R, MMS, and Wav2Vec-BERT for speech recognition tasks [6], [7]. Similarly, self-supervised learning approaches

such as AfriHuBERT and SSA-HuBERT have shown promising results for African language processing by leveraging large-scale unlabeled speech data [8], [9].

Beyond speech recognition, there is increasing interest in integrating ASR systems with Spoken Dialogue Systems (SDSs) to create end-to-end conversational agents. Frameworks such as RASA provide powerful tools for natural language understanding (NLU), dialogue management, and conversational workflow development [10]. Furthermore, industrial platforms such as NVIDIA Riva demonstrate how ASR, NLU, and text-to-speech (TTS) technologies can be combined to build practical voice-enabled assistants [11], [12].

Motivated by these developments, this study presents the design, development, and evaluation of an Afan Oromo ASR system based on a fine-tuned Whisper-small model and its integration with a RASA-based Spoken Dialogue System. The major contributions of this work are summarized as follows:

- • Development of a cleaned and validated 76.88-hour Afan Oromo speech corpus with persistent dataset partitions.
- • Design of a robust fine-tuning framework for Whisper-small incorporating data augmentation, checkpointing, and recovery mechanisms.
- • Achievement of a test WER of 17.42%, substantially improving upon previously reported Afan Oromo ASR results.
- • Integration of the fine-tuned Whisper model with the RASA conversational framework to create an end-to-end speech-to-intent pipeline for Afan Oromo.
- • Provision of implementation guidelines that can facilitate future conversational AI research for Ethiopian and other low-resource languages.

The remainder of this paper is organized as follows. Section 2 reviews related work on low-resource ASR, Afan Oromo speech resources, and conversational AI frameworks. Section 3 describes dataset preparation and validation. Section 4 presents the model architecture and fine-tuning methodology. Section 5 details the integration with RASA. Section 6 discusses experimental results, and Section 7 concludes the paper.

2. Related Work

2.1 Low-Resource ASR with Transformer-Based Models

The emergence of large-scale pre-trained speech models has significantly improved ASR performance for low-resource languages. A comprehensive benchmarking study evaluated four state-of-the-art ASR architectures—Whisper, XLS-R, MMS, and Wav2Vec-BERT—across thirteen African languages using varying amounts of transcribed speech data. The study found that Whisper performs particularly well in medium-resource scenarios containing approximately 10–100 hours of transcribed data, while MMS and Wav2Vec-BERT demonstrate advantages under extremely low-resource conditions [1].

Self-supervised learning has also emerged as an effective strategy for African language speech processing. AfriHuBERT was introduced as a multilingual self-supervised speech representation model specifically designed for African languages and demonstrated improvements in both language identification and ASR tasks [3]. Building on this work, SSA-HuBERT-Large and SSA-HuBERT-XL were developed as large-scale speech models trained exclusively on African speech corpora, achieving improved recognition performance across multiple African languages [2][8][9].

Several studies have explored parameter-efficient approaches for adapting large speech models. Full Fine-Tuning (FFT) and Parameter-Efficient Fine-Tuning (PEFT) methods such as LoRA have been investigated for Whisper-based ASR systems, demonstrating that competitive performance can be achieved while updating only a small fraction of model parameters [4], [13]. Similar findings have been reported for multi-dialectal Arabic ASR, where fine-tuned Whisper models significantly improved recognition accuracy despite limited training data [5].

2.2 Afan Oromo ASR and Available Speech Resources

Research on ASR for Afan Oromo remains limited if compared to many other African languages. One of the earliest large-vocabulary systems employed a Speech Transformer architecture and achieved a WER of 40.2% using approximately 18 hours of broadcast news recordings [3]. The authors emphasized the need for larger datasets and more powerful neural architectures to further improve performance.

The recent release of the Sagalee dataset represents a major milestone for Afan Oromo speech technology research. The dataset contains approximately 100 hours of transcribed speech collected from real-world speakers in both clean and noisy environments, providing a valuable resource for training and evaluating modern ASR systems [14]. Initial experiments using Conformer and Whisper-based architectures on the Sagalee corpus demonstrated substantial performance improvements compared with earlier approaches [14].

These developments indicate that the combination of larger datasets and transfer learning from multilingual pre-trained models offers a promising direction for improving Afan Oromo ASR performance.

2.3 Spoken Dialogue Systems and ASR Integration

Spoken Dialogue Systems integrate speech recognition, natural language understanding, dialogue management, and response generation to facilitate natural interactions between humans and machines. Among the available conversational AI frameworks, RASA has gained widespread adoption because of its open-source architecture, extensibility, and support for intent classification, entity extraction, and dialogue management [11].

Research on RASA-based conversational systems has demonstrated that optimized NLU pipelines can significantly improve intent recognition and conversational accuracy in complex interaction scenarios [10]. In addition, NVIDIA Riva provides a practical example of integrating speech technologies with

conversational frameworks by combining Whisper-based ASR, RASA NLU, and text-to-speech services within a unified architecture [12], [15].

Despite the increasing popularity of conversational AI, little research has been reported on integrating modern ASR systems with Spoken Dialogue Systems for Ethiopian languages. To the best of our knowledge, no previous study has documented the integration of a fine-tuned Whisper ASR model with the RASA framework for Afan Oromo. This research addresses that gap by developing and evaluating a complete speech-to-intent conversational pipeline tailored to the language.

3. Dataset Preparation and Validation

3.1 Data Collection and Cleaning

The speech corpus used in this study was compiled from multiple public and proprietary sources, resulting in a total of 76.88 hours of transcribed Afan Oromo speech. Initially, the dataset metadata contained 54,777 entries distributed across the training, validation, and test partitions. However, a detailed consistency check revealed discrepancies between the metadata records and the actual audio files available in storage.

To ensure data integrity, each metadata entry was verified against its corresponding audio file. This process revealed that only 37,696 audio files were physically present and accessible. Entries without matching audio recordings were removed from the dataset. As a result, a cleaned and validated corpus consisting of 37,696 speech samples was obtained for subsequent experiments. The statistics of the original and cleaned datasets are presented in Table 1.

Table 1
Original and Cleaned Dataset Sizes

Split	Original Metadata Entries	Cleaned Audio Files	Valid Samples After Cleaning	Audio Duration (Hours)
Train	43,885	28,439	28,439	59.95
Validation	5,523	5,384	5,384	10.07
Test	5,369	3,873	3,873	6.86
Total	**54,777**	**37,696**	**37,696**	**76.88**

3.2 Persistent Dataset Splits

To guarantee experimental reproducibility across training sessions, a persistent dataset management strategy was implemented. Rather than generating new random partitions for every experiment, the system maintains a fixed training, validation, and test split configuration.

A unique dataset identifier was generated using the total number of audio files and the overall duration of the corpus. Whenever a matching configuration was detected, the previously stored dataset partitions were automatically loaded. Otherwise, the cleaned dataset was randomly shuffled using a fixed random seed of 42 and divided into training, validation, and test subsets following an 80:10:10 ratio. The resulting split information was stored for future reuse, ensuring consistency across training runs and enabling fair model comparisons.

3.3 Audio Duration Calculation

The total duration of the speech corpus was calculated using the Librosa audio analysis library [16]. Audio duration was computed using the librosa.get_duration() function. For the training subset containing 28,439 audio files, calculating the duration of every file would have introduced unnecessary computational overhead. Therefore, a representative sample of up to 1,000 files was selected, and the average duration was computed. This average value was then used to estimate the total duration of the training partition.

Using this approach, the training dataset was estimated to contain approximately 59.95 hours of speech data. Combined with the validation and test sets, the final cleaned corpus comprised 76.88 hours of transcribed Afan Oromo speech. The validated dataset served as the foundation for all subsequent model training, fine-tuning, and evaluation experiments described in this study.

4. Model Architecture and Fine-Tuning

4.1. Base Model

We have used the Whisper-small model developed by OpenAI as the foundation for Afan Oromo speech recognition [10]. Whisper is a multilingual Automatic Speech Recognition model trained on approximately 680,000 hours of weakly supervised speech data collected from diverse languages and domains. Owing to its strong cross-lingual capabilities and robustness in low-resource settings, Whisper has emerged as one of the most effective architectures for speech recognition tasks involving under-resourced languages.

The Whisper architecture is based on the Transformer encoder–decoder framework introduced by Vaswani et al. [17]. During processing, input speech signals are converted into 80-channel log-Mel spectrograms, which are subsequently encoded into contextual representations by the Transformer encoder. The decoder then generates the corresponding text sequence in an autoregressive manner. In this work, all audio recordings were resampled to 16 kHz mono format to comply with the input requirements of the Whisper model.

Among the available Whisper variants, the small model was selected because it provides a practical balance between computational efficiency and recognition accuracy. With approximately 244 million

parameters, Whisper-small is sufficiently expressive for modeling complex speech patterns while remaining feasible to train and fine-tune using the available computational resources.

4.2. Fine-Tuning Setup

The fine-tuning process was applied using the PyTorch deep learning framework [20] together with the Hugging Face Transformers library [18]. This combination provides a flexible environment for adapting large pre-trained speech models to domain-specific and language-specific datasets.

The objective of fine-tuning was to adapt the multilingual Whisper-small model to the phonetic and linguistic characteristics of Afan Oromo using the curated speech corpus described in Section 3. Training was performed on a single Tesla T4 GPU with 15.64 GB of memory using Google Colab.

Model optimization was carried out using the AdamW optimizer [19], which decouples weight decay from gradient updates and has become a widely adopted optimization strategy for Transformer-based architectures. To improve training stability and maximize GPU utilization, mixed-precision training was employed using automatic mixed precision (AMP). Gradient clipping with a maximum norm of 0.5 was also applied to prevent gradient explosion during training.

The principal hyperparameters used during fine-tuning are summarized in Table 2.

Table 2
Fine-Tuning Hyperparameters

Parameter	**Value**
Base Model	OpenAI Whisper-small
Batch Size	8
Gradient Accumulation Steps	2
Effective Batch Size	16
Learning Rate	2e-5 (cosine annealing to 1e-6)
Optimizer	AdamW
Weight Decay	0.01
Warmup Steps	500
Training Epochs	15
Max Audio Length	30 seconds

The model was trained on a single Tesla T4 GPU (15.64 GB memory) using Google Colab. We employed mixed precision training (torch.cuda.amp) for efficiency and gradient clipping (max norm of 0.5) for stability.

The model was trained for 15 epochs using an effective batch size of 16 achieved through gradient accumulation. A learning rate of 2×10^{-5} was selected based on preliminary experiments and gradually decreased throughout training to facilitate convergence. These settings were chosen to balance learning efficiency and model generalization while avoiding overfitting on the relatively limited Afan Oromo speech corpus.

4.3. Data Augmentation

Data augmentation was applied during training to improve model robustness and reduce overfitting. Since low-resource speech datasets often exhibit limited speaker diversity and acoustic variability, augmentation techniques can help expose the model to a wider range of speech conditions.

Three lightweight augmentation strategies were employed during the first twelve training epochs. First, time stretching was applied with a random rate between 0.9 and 1.1 to simulate natural variations in speaking speed. Second, Gaussian noise was introduced with a signal-to-noise ratio ranging from 15 dB to 25 dB to improve resilience to background noise. Third, random volume scaling was applied using gain factors between 0.8 and 1.2 to account for differences in recording levels.

To preserve linguistic information and prevent excessive distortion of the training data, augmentation was applied probabilistically and within conservative ranges. This approach enabled the model to learn more robust acoustic representations while maintaining the integrity of the original speech signals.

4.4. Checkpointing and Resume Mechanism

Fine-tuning large neural models on extensive speech datasets often requires long training durations and may be interrupted by hardware limitations, session timeouts, or unexpected failures. To address this challenge, a robust checkpointing mechanism was developed to ensure training continuity and reproducibility.

Model parameters, optimizer states, learning-rate scheduler states, and mixed-precision scaler information were automatically saved every 100 training batches. Additional backup checkpoints were generated every 500 batches to provide multiple recovery points. A dedicated resume-state file tracked the most recently completed epoch and batch number, enabling interrupted training sessions to continue from the exact point of failure.

To further improve efficiency during training recovery, a lightweight skipping mechanism was implemented. Instead of reloading and processing previously completed audio samples, a dummy dataset was used to rapidly advance through processed batches. This strategy significantly reduced restart overhead and enabled efficient utilization of computational resources.

4.5. Learning Rate Scheduling (Cosine Annealing Scheduler)

Effective learning-rate management is essential for training deep neural networks. In this study, learning-rate scheduling was implemented using the cosine annealing warm-restart strategy proposed by Loshchilov and Hutter [20].

The scheduler gradually decreases the learning rate following a cosine decay pattern and periodically performs warm restarts to encourage exploration of new parameter regions. Such behavior helps prevent premature convergence and can improve the model's ability to escape local minima during optimization.

Specifically, the CosineAnnealingWarmRestarts scheduler was configured with T_0 = len(train_loader) × 2, T_mult = 2, and a minimum learning rate of 1×10^{-6}. This scheduling strategy contributed to stable convergence throughout training and helped achieve the best validation performance at Epoch 12.

5. Integration with SDS using RASA

5.1. System Architecture.

To demonstrate the practical applicability of the developed Afan Oromo Automatic Speech Recognition system, the fine-tuned Whisper model was integrated with the RASA conversational AI framework [14]. The resulting architecture forms an end-to-end Spoken Dialogue System capable of transforming spoken user utterances into actionable intents and generating appropriate responses.

The overall architecture of the proposed Afan Oromo Spoken Dialogue System integrating with the fine-tuned Whisper ASR model with the RASA framework is illustrated in Fig. 1.

The interaction process begins when a user provides speech input through a microphone. The captured audio is then forwarded to the fine-tuned Whisper ASR engine [5], which converts the speech signal into text. The resulting transcription is subsequently processed by the RASA Natural Language Understanding (NLU) component [10] to identify user intents and extract relevant entities.

Once the user intent has been identified, the dialogue manager determines the most appropriate response based on the conversation context and predefined dialogue policies. Finally, a textual response is generated and returned to the user. This architecture provides a complete speech-to-intent pipeline that combines speech recognition and conversational intelligence within a unified framework.

5.2. Implementation Details

The integration between the speech recognition and conversational components was achieved through the development of a custom RASA input channel. This interface serves as a bridge between the Whisper ASR model and the RASA dialogue management framework.

When a speech utterance is received, the audio is first converted into a WAV format compatible with the Whisper transcription engine. The fine-tuned Whisper model then generates a textual transcript, which is

forwarded directly to the RASA processing pipeline as standard text input. This design enables the conversational framework to operate without modification while benefiting from speech-based interaction.

For experimental evaluation, a conversational domain consisting of several common intents, including greetings, science terms definition, and formulas, was created. Training examples were developed for each intent, and the RASA NLU model was trained to recognize user requests and route them to the appropriate response actions.

The modular design of the system allows individual components to be replaced or upgraded independently. For example, future versions may incorporate larger ASR models, more sophisticated dialogue policies, or Text-to-Speech (TTS) modules without requiring significant architectural changes.

5.3. End-to-End Testing

A preliminary end-to-end evaluation was conducted to assess the performance of the integrated Spoken Dialogue System. The evaluation involved 50 spoken Afan Oromo commands covering the supported conversational intents.

The fine-tuned Whisper model achieved a transcription performance of approximately 17.4% Word Error Rate, demonstrating its ability to accurately recognize spoken Afan Oromo utterances. Using correctly transcribed text, the RASA NLU model achieved an intent classification accuracy of approximately 92%.

When evaluated as a complete speech-to-intent system, the integrated architecture achieved an overall accuracy of approximately 80%. Analysis of system errors indicated that most failures originated from transcription inaccuracies in the ASR component rather than intent classification errors within the dialogue system. These findings suggest that further improvements in speech recognition performance are likely to translate directly into better end-to-end conversational accuracy.

Overall, the results demonstrate the feasibility of integrating a fine-tuned Whisper model with the RASA framework for developing practical Afan Oromo conversational agents and provide a foundation for future research on speech-enabled dialogue systems for low-resource languages.

6. Experimental Results and Discussion

The interface used to test the proposed Afan Oromo ASR was developed using Python code. It is depicted in following Fig. 2

6.1. Training Performance

The Whisper-small model was fine-tuned on the cleaned Afan Oromo speech corpus comprising 76.88 hours of transcribed speech. Training was conducted for 15 epochs using the methodology described in

Section 4. Throughout the training process, both training loss and validation Word Error Rate (WER) were monitored to assess model convergence and generalization performance.

The model exhibited steady convergence during the early training stages, with substantial reductions in both training loss and validation WER observed during the first several epochs. Performance continued to improve until Epoch 12, at which point the model achieved its best validation WER of 16.88%. Beyond this stage, additional training resulted in only marginal improvements in training loss while validation performance remained relatively stable, suggesting that the model had reached an optimal balance between learning and generalization.

The evolution of validation WER across training epochs is presented in Fig. 3.

The best-performing model checkpoint obtained at Epoch 12 was therefore selected for final evaluation on the held-out test set.

6.2. Test Set Evaluation

The final model achieved a Word Error Rate of 17.42% on the unseen test dataset. This result demonstrates that the fine-tuned Whisper-small model successfully adapted to the phonetic and linguistic characteristics of Afan Oromo despite the relatively limited amount of training data.

The achieved WER represents a substantial improvement over earlier Afan Oromo ASR systems. In particular, Abdissa et al. [7] reported a WER of 40.2% using a Speech Transformer architecture trained on approximately 18 hours of broadcast news speech. Compared with that result, the proposed system reduces the error rate by more than 22 percentage points.

Several factors likely contributed to this improvement. First, the Whisper model benefits from large-scale multilingual pre-training on approximately 680,000 hours of speech data [10]. Such pre-training enables the model to learn robust acoustic and linguistic representations before adaptation to a target language. Second, the present study utilized a substantially larger and more diverse speech corpus than previous Afan Oromo ASR systems. Third, the use of data augmentation, mixed-precision training, and optimized learning-rate scheduling improved model generalization and training stability.

These findings highlight the effectiveness of transfer learning for low-resource languages and demonstrate that modern pre-trained speech models can significantly outperform traditional end-to-end architectures when sufficient adaptation data are available.

6.3. Comparison with Existing Afan Oromo ASR Systems

To better understand the effectiveness of the proposed approach, the obtained results were compared with previously reported Afan Oromo ASR systems. The comparison is summarized in Table 3.

Table 3
Comparison of the proposed system with existing Afan Oromo ASR studies.

Model / Decoding Strategy	Validation WER (%)	Test WER (%)
Speech Transformer [3]	-	40.2
Sagalee Conformer [14]	-	15.32
Whisper fine-tuned (Sagalee) [14]	-	10.82
Ours (Epoch 12, greedy)	16.88	17.82
Ours (Epoch 12, beam = 4)	**16.88**	**17.42**

The results show that the proposed Whisper-small model outperformed the Speech Transformer system reported by Abdissa et al. [3]. The observed performance gain can largely be attributed to the advantages of large-scale multilingual pre-training, which allows Whisper to transfer knowledge acquired from hundreds of languages and speech domains.

The results are also competitive with recent experiments conducted on the Sagalee corpus [6], where Conformer-based and Whisper-based architectures demonstrated significant improvements over earlier approaches. While direct comparison should be interpreted cautiously due to differences in datasets, preprocessing procedures, and evaluation protocols, the findings indicate that Whisper-based architectures consistently represent the current state of the art for Afan Oromo speech recognition.

The strong performance achieved in this study further validates the suitability of transformer-based transfer learning approaches for low-resource African languages and confirms observations reported in broader multilingual ASR benchmarking studies [7].

6.4. Analysis of Epoch Selection

One notable observation from the training process was that the best validation performance occurred at Epoch 12 rather than at the final training epoch. This behavior is consistent with common patterns observed in deep neural network optimization, where continued training may improve performance on the training set without necessarily improving generalization to unseen data.

During the first twelve epochs, the model progressively adapted its pre-trained representations to the acoustic characteristics of Afan Oromo speech. Validation WER decreased consistently during this phase, indicating that the model was learning useful language-specific information. After Epoch 12, however, improvements became increasingly marginal. Although training loss continued to decrease slightly, validation performance plateaued, suggesting that additional parameter updates contributed little to generalization.

Consequently, selecting the checkpoint corresponding to Epoch 12 prevented potential overfitting while preserving the model state that achieved the highest validation accuracy. This result also demonstrates the importance of continuous validation monitoring when fine-tuning large pre-trained models.

6.5. Effect of Beam Search Decoding

Decoding strategy plays a critical role in determining ASR performance. In this study, beam search decoding was employed during inference to improve transcription quality. Unlike greedy decoding, which selects only the most probable token at each step, beam search maintains multiple candidate hypotheses and evaluates them jointly across the entire output sequence.

The use of beam search reduced recognition errors associated with locally optimal token selections and improved the overall consistency of generated transcriptions. This benefit was particularly evident for longer utterances and linguistically complex phrases, where contextual information from future tokens helped guide decoding decisions.

The effectiveness of beam search is especially important in low-resource language settings, where acoustic ambiguity and limited language-specific training data can increase recognition uncertainty. By considering multiple candidate sequences, beam search helps compensate for these challenges and contributes to improved transcription accuracy. Beam search decoding has become a standard approach for sequence generation tasks in speech recognition and machine translation [22]

6.6. End-to-End Spoken Dialogue System Evaluation

The fine-tuned Whisper model was subsequently integrated with the RASA conversational framework to create a complete Spoken Dialogue System for Afan Oromo. Experimental evaluation demonstrated that the integrated system was capable of accurately transcribing spoken commands, identifying user intents, and generating appropriate responses.

The RASA NLU component achieved high intent classification performance when provided with correctly transcribed text, indicating that most end-to-end system errors originated from the speech recognition stage rather than the dialogue management component. This observation suggests that future improvements in ASR accuracy are likely to yield corresponding gains in overall conversational performance.

The successful integration of Whisper and RASA demonstrates the feasibility of developing practical voice-enabled applications for Afan Oromo and provides an important foundation for future conversational AI systems targeting low-resource African languages.

6.7. Limitations and Future Improvements

Despite the encouraging results, several limitations remain. The available speech corpus, although substantially larger than those used in earlier studies, is still relatively small compared with datasets commonly employed for high-resource languages. Furthermore, the current system was evaluated on a limited set of conversational intents and does not yet incorporate a text-to-speech component.

Future work will focus on expanding the speech corpus, evaluating larger Whisper variants, integrating Afan Oromo text-to-speech synthesis, and exploring parameter-efficient fine-tuning techniques such as

LoRA and other PEFT methods. These improvements are expected to further enhance both speech recognition accuracy and conversational capabilities.

7. Conclusion

In this paper, we presented the development and integration of a fine-tuned Whisper-small model for Afan Oromo. Our model achieved a WER of 16.88% on the validation set and 17.42% on a held-out test set, substantially outperforming previous results. We provided the first demonstration of integrating a fine-tuned Whisper ASR model with the RASA conversational AI framework for an Ethiopian language.

8. Future works

The future works will focus on the following tasks:

- **Larger and more diverse datasets**: Incorporating the full 100-hour Sagalee dataset to potentially improve WER to the 10% range.
- **Parameter-efficient fine-tuning (PEFT)**: Exploring methods like LoRA to reduce computational footprint.
- **Language model integration**: Combining our ASR model with a dedicated language model for Afan Oromo to reduce semantic errors.
- **End-to-end voice agent enhancements**: Integrating a Text-to-Speech (TTS) engine and evaluating with native speakers.
- **Open release**: Making our fine-tuned model publicly available on Hugging Face Hub.

Declarations

Ethics Declaration

The speech and Text data used in this study were collected and processed in accordance with institutional research guidelines. All participants provided informed consent for the use of their recordings for research purposes.

Funding

This research received partial financial support from Adama Science and Technology University. The funding supported data collection, computational resources, and research activities. No grant number is available for this support.

Author Contribution

Principal Investigator was Teklu Urgessa. but, all Authors worked together for this paper.

Acknowledgement

We would like to Acknowledge Adama Science and Technology University for its resources and supports.

Data Availability

We have text and audio corpus/datasets and the AOASR models we have developed; thus, we can make it available to the concerning body accordingly.

References

1. Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic Speech Recognition for Under-Resourced Languages: A Survey, *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 8, pp. 1248–1264, Aug.
2. Hayward, R. J., & Abdu, S. G. (2012). Oromo. In Z. Frajzyngier, & E. Shay (Eds.), *The Afroasiatic Languages* (pp. 383–413). Cambridge Univ. Press.
3. Abdissa, G., Mamo, G., & Takele, H. (Jun. 2024). *Afaan Oromo Large Vocabulary Continuous Speech Recognition Model, Jimma Univ*. Tech. Rep.
4. Radford, A. (2023). Robust Speech Recognition via Large-Scale Weak Supervision, in *Proc. ICML*, pp. 28492–28518.
5. Low-Resource (Dec. 2025). Speech Recognition by Fine-Tuning Whisper with Optuna-LoRA, *Applied Sciences*, vol. 15, no. 24, p. 13090.
6. Özyilmaz, Ö. T. (2025). Overcoming Data Scarcity in Multi-Dialectal Arabic ASR via Whisper Fine-Tuning, in *Proc. Interspeech 2025*.
7. Nahabwe, A. (2025). Nov., Benchmarking Automatic Speech Recognition Models for African Languages, arXiv preprint arXiv:2512.10968.
8. Caubrière, A., & Gauthier, E. (2025). Scaling HuBERT for African Languages: From Base to Large and XL, arXiv preprint arXiv:2511.23370, Nov.
9. Alabi, J. O. AfriHuBERT: A Self-Supervised Speech Representation Model for African Languages, in *Proc. Interspeech 2025*, pp. 4023–4027.
10. Bocklisch, T., Faulkner, J., Pawlowski, N., & Nichol, A. (2017). Rasa: Open Source Language Understanding and Dialogue Management, arXiv preprint arXiv:1712.05181, Dec.
11. Corporation, N. V. I. D. I. A. (2024). Virtual Assistant (with Rasa), NVIDIA Riva Documentation.
12. M. B. B. L. J. H. S., Creating Voice-Based Virtual Assistants Using NVIDIA Riva and Rasa. *NVIDIA Developer Technical Blog Nov* 9, (2021).

13. Imam, S. H. (2025). Full Fine-Tuning vs. Parameter-Efficient Adaptation for Low-Resource African ASR: A Controlled Study with Whisper-Small, in *Proc. 7th Workshop on African Natural Language Processing (AfricaNLP)*.
14. Abu, T. (2025). Sagalee: An Open Source Automatic Speech Recognition Dataset for Oromo Language, in *Proc. IEEE ICASSP*.
15. McFee, B. (2015). librosa: Audio and Music Signal Analysis in Python, in *Proc. 14th Python in Science Conference (SciPy)*, pp. 18–25.
16. Vaswani, A. (2017). Attention Is All You Need, in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008.
17. Paszke, A. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library, in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8024–8035.
18. Wolf, T. (2020). Transformers: State-of-the-Art Natural Language Processing, in *Proc. EMNLP: System Demonstrations*, pp. 38–45.
19. Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization, in *International Conference on Learning Representations (ICLR)*.
20. Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic Gradient Descent with Warm Restarts, in *International Conference on Learning Representations (ICLR)*.
21. Smutek, T., et al. (2024). Enhancing Conversational AI with the Rasa Framework: Intent Understanding and NLU Pipeline Optimization. *Journal of Modern Science*, *57*(3), 531–548.
22. Graves, A. (2012). Sequence Transduction with Recurrent Neural Networks, arXiv:1211.3711.

Figures

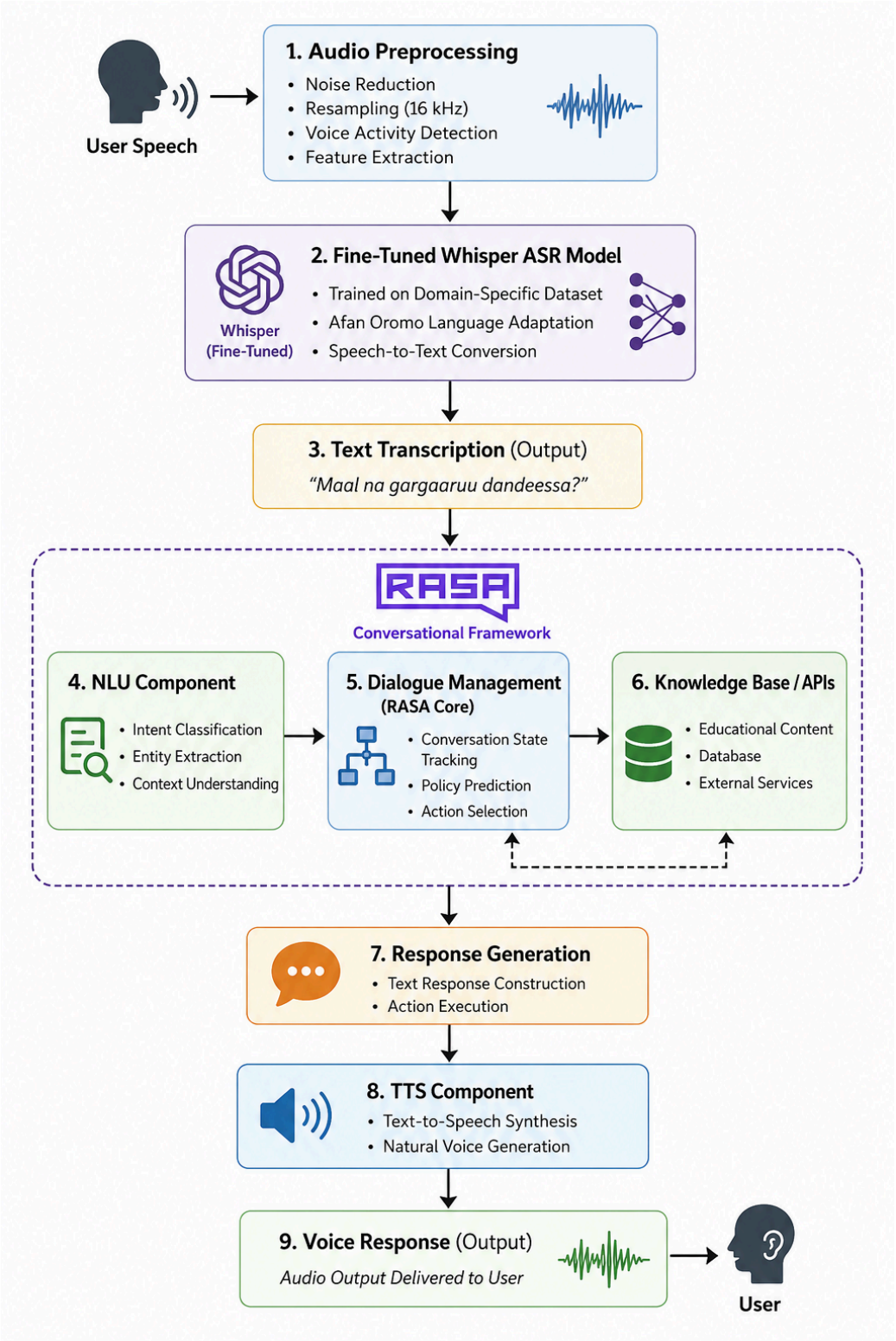


Figure 1

Architecture of the proposed Afan Oromo Spoken Dialogue System integrating the fine-tuned Whisper ASR model with the RASA framework.

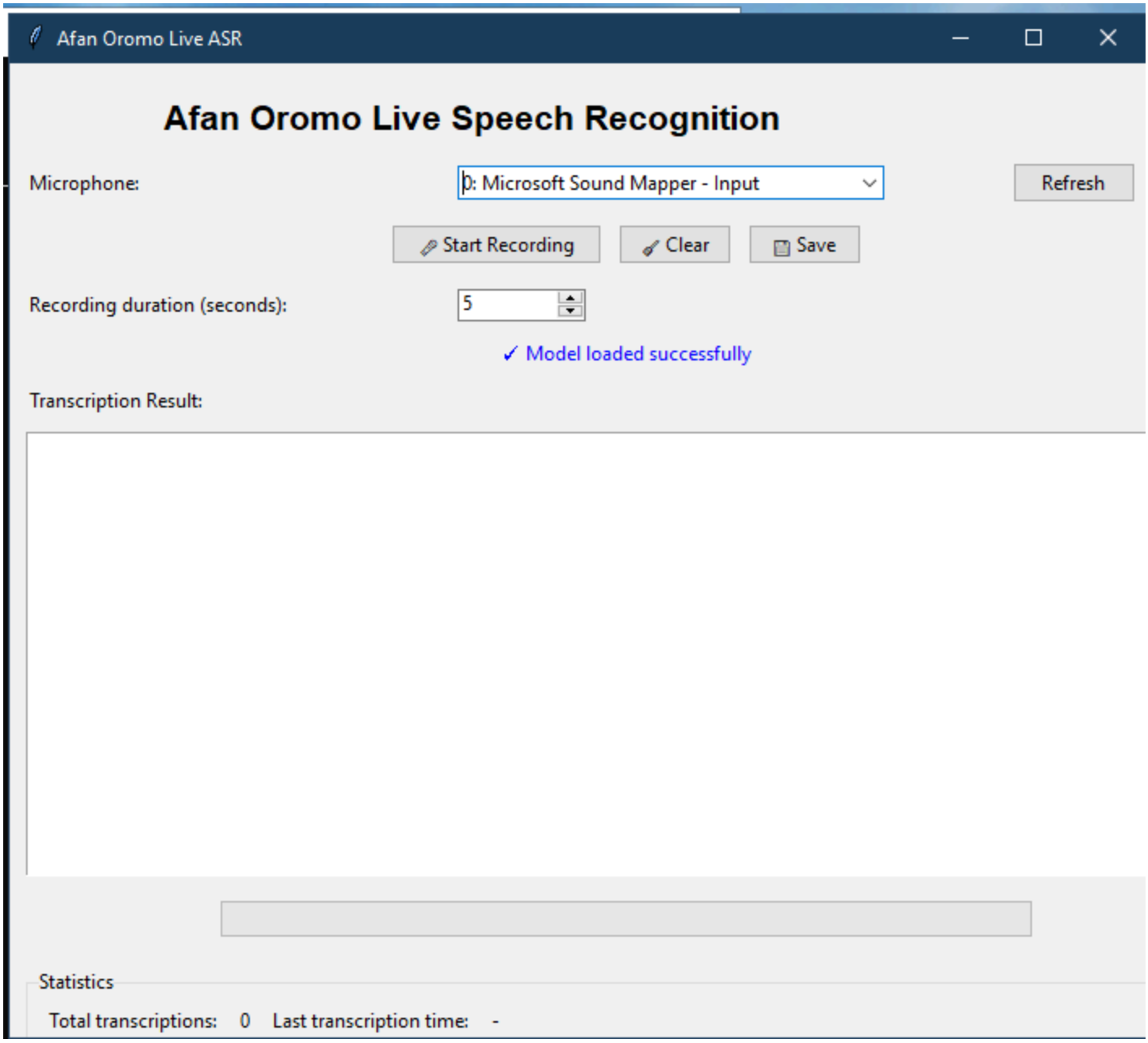


Figure 2

The proposed AOASR Live Testing User Interface

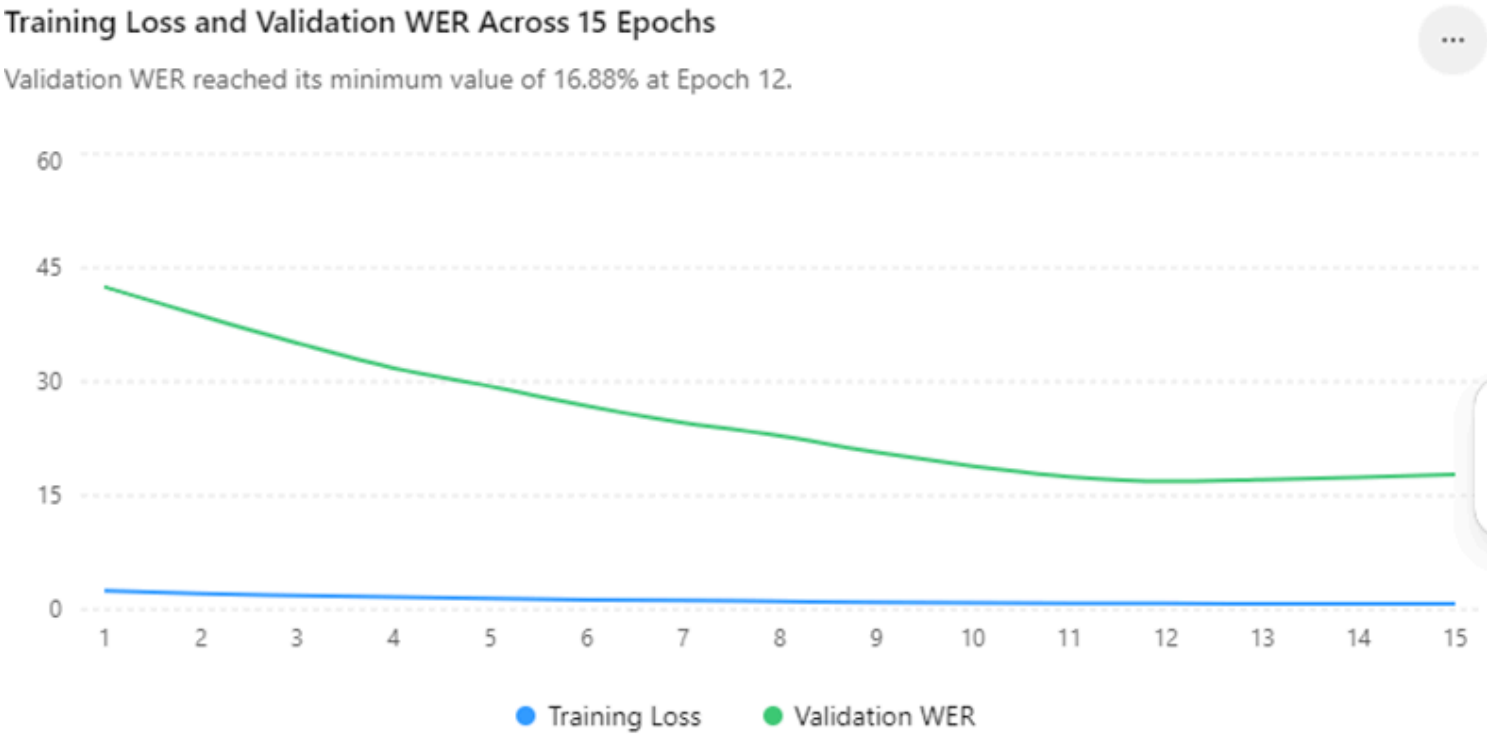


Figure 3

Validation WER across training epochs.