A Unified End-to-End Approach for Amharic Speech Recognition Using Clean, Noisy, and Synthetic Data

Yohannes Ayana Ejigu
Development Team, AI research
Amplitude Ventures AS
Stavanger, Norway
yohannesayana10@gmail.com

Tesfa Tegegne Asfaw
Faculty of Computing
Bahir Dar University
Bahir Dar, Ethiopia
tesfat@gmail.com

Surafel Amsalu Tadesse
Development Team, AI research
Amplitude Ventures AS
Stavanger, Norway
surafelamsalu2013@gmail.com

Anteneh Yehalem Tegegne
Department of Data Science
Bahir DarUniversity
Bahir Dar, Ethiopia
anteneh.yehalem@bdu.edu.et

***Abstract*—Automatic Speech Recognition for low-resource languages such as Amharic faces challenges due to limited high-quality data and background noise. This study examines how different types of training data, including clean recordings, noisy recordings, and synthetically augmented data, affect the performance and robustness of an Amharic speech recognition system. The experiments use 155 hours of speech, including 110 hours of clean data from the Andreas Nürnberger Data and Knowledge Engineering Group and 45 hours of real-world noisy recordings. Additional synthetic data were created using noise injection, speed perturbation, and SpecAugment techniques, resulting in a total of 575 hours of speech data. A convolutional neural network with bidirectional gated recurrent units and Connectionist Temporal Classification was trained on four conditions: clean data, noisy data, combined data, and augmented data. The results show that models trained on combined and augmented datasets outperform models trained on a single dataset, achieving a word error rate of 5.49 percent under mixed conditions, a relative improvement of 21.5 percent. These findings highlight the importance of data diversity and augmentation in developing robust speech recognition systems for low-resource languages. Future work will explore the use of visual information, such as lip movements to further improve recognition accuracy in challenging environments.**

Keywords—Amharic ASR, noise robustness, data augmentation, end-to-end models, low-resource languages

I. INTRODUCTION

Automatic Speech Recognition (ASR) has become an essential part of modern computing, powering tools such as transcription software, virtual assistants, and interactive voice response systems [1], [2]. Recent progress in deep-learning methods has led to remarkable accuracy for high-resource languages such as English and Mandarin [3], [4]. However, low-resource languages like Amharic still face serious challenges, especially when speech is recorded in noisy or uncontrolled environments [5], [6].

Earlier ASR systems relied on hybrid designs that combined acoustic, pronunciation, and language models [7], [8]. While effective, these pipelines required hand-crafted linguistic rules and expert supervision [9].

End-to-end models simplified the process by mapping acoustic features directly to text, using architectures such as Connectionist Temporal Classification (CTC) [10], encoder–decoder models [11],[21], and Conformer networks [12]. Despite these advances, robustness to background noise and data imbalance remains a limiting factor for many low-resource languages [17], [18].

For Amharic--a Semitic language spoken by more than sixty million people--the majority of existing ASR systems have been trained on clean, studio-quality data [19], [16]. These systems often fail in real-world conditions, where background sounds such as traffic, voices, or electronic interference can distort the signal [17]. Addressing this problem requires not only architectural improvements but also careful attention to the type and diversity of data used for training [18].

Although several studies have introduced new neural architectures for Amharic ASR, fewer have examined how the composition of data influences model robustness. Recent findings in multilingual and low-resource ASR research suggest that data-centric approaches can be as powerful as model-centric ones [19], [20].

Data augmentation techniques--such as SpecAugment [21], noise injection [22], and speed perturbation [22]--expand existing corpora and help models generalize to unseen conditions without costly manual collection. These strategies have been widely tested on large languages but rarely explored in depth for Amharic. Building on this perspective, the present study evaluates how training data quality and diversity affect the performance of an Amharic end-to-end ASR model. Instead of developing a new network architecture, we adopt a stable CNN–BiGRU–CTC baseline [24] and train it under four data conditions: clean, noisy, combined, and synthetically augmented.

The clean corpus (110 hours) originates from the Andreas Nürnberger Data and Knowledge Engineering Group [17], while the noisy dataset (45 hours) consists of natural Amharic speech gathered in public environments and previously used for model-development research [18]. Synthetic variations were created through well-known

augmentation methods to simulate realistic speaking and acoustic conditions.

This work provides an empirical view of how dataset composition, rather than architectural complexity, shapes model robustness for low-resource languages. The insights gained from Amharic may also support similar efforts across other African and morphologically rich languages where labeled speech resources remain limited.

A. Research Questions

This study aims to explore how different types of training data influence the robustness of Amharic end-to-end speech recognition systems. The work focuses on understanding data effects rather than introducing a new model. Accordingly, the research is guided by the following key questions:

i. How does the performance of the Amharic ASR model vary when trained on clean, noisy, and combined datasets?
ii. Which training condition provides the most stable and accurate results under varying noise levels?
iii. How does the inclusion of synthetic (augmented) data influence the model's ability to generalize to unseen speech and noise conditions?

These questions aim to clarify the relative contributions of dataset quality, diversity, and augmentation to overall ASR robustness.

II. DATA

We yes This study utilized a total of 155 hours of Amharic speech data collected from clean and noisy environments, later expanded to an effective 575 hours through systematic data augmentation. The data were designed to represent diverse acoustic conditions and speaker characteristics for developing a noise-robust ASR model.

A. Clean Dataset

The clean dataset, totaling 110 hours, was obtained from the Andreas Nürnberger – Data and Knowledge Engineering Group. These recordings were produced in controlled acoustic environments with minimal background interference. The corpus includes both male and female speakers of various ages and regional accents, ensuring diversity in speech rate and pronunciation style. The recordings are of high quality, sampled at 16 kHz, and serve as the baseline for evaluating clean-condition ASR performance.

B. Noisy Dataset

The noisy dataset, totaling 45 hours, was collected in the Sidama region of Ethiopia, where Amharic is widely used as a second language. This region was intentionally chosen for its linguistic diversity and mixture of urban and rural environments. Speech was recorded from 50 speakers, each reading 400 distinct sentences, producing 20,000 utterances. Each audio clip ranged from 4 to 20 seconds.

Recordings were made in natural noisy locations—cafés, streets, parks, and marketplaces—where background sounds such as traffic noise, wind, birdsong, and human conversation were present. Equal participation of male and female speakers aged 18 – 50 years ensured demographic balance. All recordings were manually transcribed into Amharic script to provide accurate ground truth for model training.

C. Combined and Synthetic Data

To improve model robustness and generalization, the clean (110 h) and noisy (45 h) datasets were merged, forming a 155-hour combined corpus. Three augmentation techniques were then applied to expand the training diversity:

Speed Perturbation: Each utterance was resampled at ±10% speed variations, creating two additional versions per sample.

→ Adds 2× the original data (310 h new), simulating different speaking rates.

Noise Injection: Background noise was added to each clean recording at 5 – 25 dB SNR, producing one additional noisy version of the clean corpus.

→ Adds 110 h of new data, capturing varied interference levels.

SpecAugment: Applied on-the-fly during training by masking time and frequency regions in spectrograms, improving robustness without increasing stored data size.

After augmentation, the effective training data totaled approximately 575 hours. This expanded dataset offers richer acoustic and temporal diversity, helping prevent overfitting and improving performance in real-world conditions.

D. Feature Extraction

All audio signals were transformed into spectrograms using the Short-Time Fourier Transform (STFT). Each waveform was divided into overlapping frames (25 ms window, 10 ms stride), and a Fourier transform was applied to extract the frequency components of each frame. The resulting magnitude spectrograms provide a two-dimensional time–frequency representation, which preserves both spectral and temporal dynamics of Amharic speech. These spectrograms were normalized and used as input features for model training.

The dataset was split into training, validation, and test subsets in an 80/10/10 ratio. Transcriptions are paired with audio filenames, as illustrated in Fig. 1. For feature extraction, audio was converted into spectrograms using Short-Time Fourier Transform (STFT). These spectrograms serve as input to the neural network models. The complete dataset is openly available at:

https://figshare.com/articles/dataset/Yohannes_A_Ejigu_Amharic_ASR_Dataset_zip/24959727

The transcribed text paired with the file name of the audio is presented in Fig. 1 below.

	normalized_transcription	
0	የተ ለቀቁት ምርኮኞች በ አካባቢያቸው ለለማዊ ኑሮ እንዲ ኖሩ የ ትራንስፖ...	tr_2_tr010
1	አንቶኔሊ በ አጼ ምንሊክ ፊት የ ፈጸመው ድፍረት በ ኢጣሊያ ን ምከር ቤት...	tr_9_tr01009 - Co
2	ላቁኦተ ትምህርት ቤት መንገድ ና ሆስፒታል ተ ገንብቷል	tr_14_tr010

Fig. 1 presents the top 3 rows of the data

III. SYSTEM DESCRIPTION

The proposed system is an end-to-end Amharic Automatic Speech Recognition (ASR) model designed to maintain high performance across varying acoustic conditions. It integrates convolutional, recurrent, and alignment-free decoding layers,

forming a hybrid architecture optimized for low-resource and noise-prone environments.

Unlike prior works that focused primarily on network depth or feature engineering, this research emphasizes the interaction between data diversity and model robustness, evaluating performance under clean, noisy, and augmented training conditions.

A. Model Architecture

The model follows a CNN–RNN–CTC framework widely adopted in modern speech recognition systems [17], [18]. The convolutional front end extracts spatial and spectral cues from spectrograms, the recurrent layers capture sequential and contextual dependencies, and the Connectionist Temporal Classification (CTC) layer aligns the predicted and reference sequences without requiring frame-level annotations.

1) Input Representation:

Each audio sample is converted into a time–frequency spectrogram through the Short-Time Fourier Transform (STFT). This representation encodes both temporal and frequency variations essential for recognizing Amharic's complex syllabic structure. Spectrograms are normalized to ensure consistent dynamic ranges across speakers and recording conditions.

2) Convolutional Feature Extraction:

Two-dimensional convolutional layers are used to capture local acoustic features, such as formant transitions and energy contours. Each convolutional block is followed by batch normalization and ReLU activation, improving training stability and reducing sensitivity to amplitude variations. The convolutional stage provides noise tolerance by learning invariant frequency–time patterns.

3) Temporal Modeling (BiGRU):

The convolutional features are reshaped and fed into a stack of Bidirectional Gated Recurrent Unit (BiGRU) layers, each with 512 hidden units. Bidirectionality enables the model to process information from both forward and backward time directions, improving its ability to model long contextual dependencies.

4) Dense Projection and CTC Decoding:

The recurrent output is passed through a fully connected projection layer that maps the hidden states to character probabilities. A CTC decoding layer aligns these predictions with the ground-truth transcripts by introducing a "blank" token between consecutive labels. This allows the model to learn flexible mappings between variable-length audio and text sequences without explicit segmentation.

The overall architecture is illustrated in Fig. 2, showing the full pipeline from spectrogram input to character-level transcription.

The overall architecture is illustrated in Fig. 2, showing the flow from spectrogram input to character-level output.

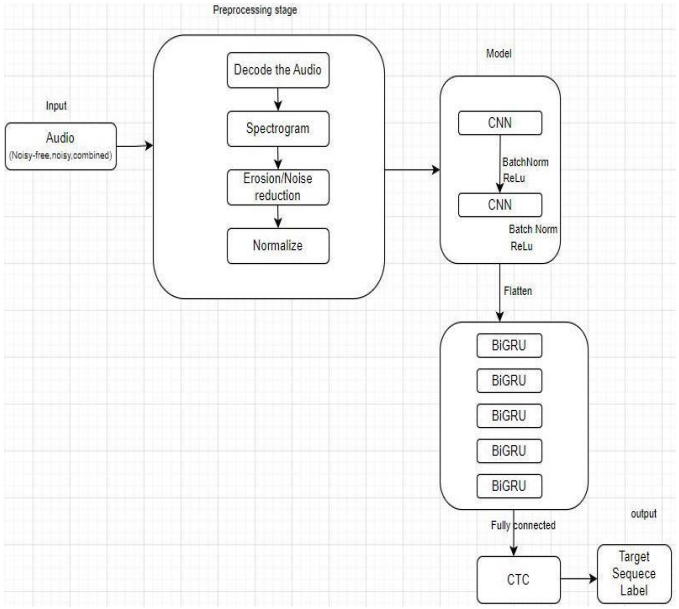

Fig2. Proposed Architecture of the model [17,18]

A. Model Implementation

The The Amharic end-to-end ASR model was implemented using the TensorFlow/Keras framework. The design follows a CNN–BiGRU–CTC structure, chosen for its simplicity and robustness in handling speech variations across different noise levels and accents. The model architecture is shown in Fig. 2.

Input Representation and Preprocessing: Audio recordings from all dataset conditions--clean, noisy, combined, and augmented--were first converted into spectrograms using the Short-Time Fourier Transform (STFT). This representation captures both temporal and spectral information, which is essential for distinguishing speech features in low-resource languages like Amharic. Each spectrogram was normalized and reshaped into 3D tensors before being passed to the neural network to ensure consistent input dimensions.

Feature Extraction via CNN: The spectrograms were fed into two 2D convolutional layers, each followed by batch normalization and ReLU activation. These layers capture local acoustic patterns, such as formants and harmonics, while providing resilience to background noise and speaker variations. The convolutional outputs were then flattened along the frequency dimension to prepare them for sequence modeling.

Temporal Modeling with BiGRU: The flattened features were passed through a stack of Bidirectional GRU (BiGRU) layers, each containing 512 hidden units in both forward and backward directions. This configuration allows the model to process contextual information from both past and future frames, which is particularly valuable for Amharic's morphologically rich structure. A dropout rate of 0.5 was applied to mitigate overfitting. For comparative analysis, a BiLSTM variant was also trained under identical settings to evaluate trade-offs between accuracy and computational efficiency.

Dense and Output Layers: Outputs from the recurrent layers were passed through a fully connected dense layer to project the learned representations into the character probability space. Finally, a softmax layer produced the output

distributions, including the "blank" token required for Connectionist Temporal Classification (CTC) training.

Sequence Alignment and Decoding: The model was optimized using the CTC loss function, which enables alignment-free training between input speech and text transcriptions. During inference, decoding was performed using both beam search (for accuracy) and greedy decoding (for speed) to convert predicted probability sequences into readable text. errors.

B. Model Compilation

Training was carried out with the Adam optimizer (learning rate = 1×10^{-4}). The combination of CTC loss and Adam optimization ensured stable convergence for variable-length Amharic speech data. The system was trained for multiple epochs, with early stopping based on validation loss to prevent overfitting.

Let y be the ground truth label sequence, x be the input sequence, and y* be the predicted label sequence. The CTC loss function is defined as:

$$\text{CTC Loss}(x,y) = -\log \sum_{\pi \in \text{Align}(x,y)} \prod_{t=1}^{T} P(\pi_t|x) \qquad (1)$$

where:

T is the length of the input sequence x,

π represents a possible alignment between x and y

$\text{Align}(x,y)$ is the set of all valid alignments between x and y,

π_t is the label assigned to time step t in the alignment π, and

$P(\pi_t|x)$ is the probability assigned to label π_t at time step t by the neural network model. [17,18]

C. Evaluation Metric

Model performance was assessed using the Word Error Rate (WER), a widely adopted measure in ASR research. WER is defined as the proportion of substitutions, deletions, and insertions required to transform the system output into the reference transcription. It is computed as

$$\text{WER} = \frac{S+D+I}{N} \qquad (2)$$

where S, D, and I denote the number of substitutions, deletions, and insertions, respectively, and N is the total number of words in the reference. Lower WER values correspond to higher transcription accuracy, making the metric useful for both model evaluation and hyperparameter tuning.

IV. Results

The experiments were designed to examine the impact of different training conditions--clean, noisy, combined, and synthetically augmented--on the robustness of the Amharic ASR model. Each model was trained under identical settings using the CNN–BiGRU–CTC architecture and evaluated on both clean and noisy test sets.

A. Dataset-Based Performance Comparison

To address the first research question, three baseline models were trained separately on clean, noisy, and combined datasets. The results show that the model trained solely on clean speech achieved the lowest error rate in quiet test conditions but performed poorly in noisy environments. Conversely, the model trained only on noisy data generalized well to unseen noise but exhibited reduced accuracy on clean speech. The combined dataset offered a balanced trade-off, performing consistently across both conditions.

Table I presents performance comparison in WER

Training Dataset	Test Condition	WER (%)
Clean only (110 h)	clean	7.1
Clean only	Noisy	14.8
Noisy only (45 h)	Noisy	10.25
Noisy	Clean	13.4
combined(155 hrs)	clean	6.88
combined	Noisy	11.7

These findings (in table I) demonstrate that combining clean and noisy data during training significantly improves model generalization without introducing overfitting to specific noise types.

B. Effect of Synthetic Augmentation

To investigate the third research question, three standard augmentation techniques--noise injection, speed perturbation, and SpecAugment—were applied to the combined dataset, producing a total of 575 hours of effective training data.

The augmented model achieved a Word Error Rate (WER) of 5.49% on mixed test conditions, outperforming all baselines. This indicates that synthetic data plays a substantial role in enhancing robustness, especially when authentic noisy data is limited. Among the augmentation techniques, SpecAugment provided the most noticeable gain by helping the network become invariant to time–frequency distortions, while speed perturbation contributed to improved speaker and tempo diversity. The effect of synthethic augmentation is seen in Table 2 below.

Table2. presents effects of synthetic augmentation

Training Data	Augmentation applied	WER (%)	Relative improvement
Combined (155 h)	None	6.88	-
Combined + Speed Perturbation	Yes	6.21	9.74%
Combined + Noise Injection	Yes	6.18	10.17%
Combined + All Techniques (575 h total)	Yes	5.49	21.51%

V. Discussions

The experimental results confirm that data diversity significantly enhances ASR robustness. Models trained solely on clean data performed best in noise-free conditions (7.1% WER) but degraded under noisy settings (14.8%), while noisy-only training improved noise tolerance but reduced accuracy on clean inputs. The combined dataset achieved balanced performance (6.88% WER clean, 11.7% noisy), aligning with prior studies that emphasize mixed-domain exposure for improved generalization.

Further, synthetic augmentation yielded substantial gains. Using noise injection, speed perturbation, and SpecAugment increased the effective dataset size to 575 hours and reduced WER to 5.4%, representing a 21.5% relative improvement. This confirms that data augmentation can effectively compensate for limited real recordings in low-resource settings [3], [4]. The BiGRU-based model demonstrated stable convergence and superior efficiency over other alternatives, consistent with earlier ASR findings . Overall, the combination of authentic and augmented data proved effective for enhancing robustness and generalization in Amharic speech recognition.

VI. CONCLUSIONS

This study presented an end-to-end Amharic Automatic Speech Recognition (ASR) framework using a CNN–BiGRU–CTC architecture. The system was rigorously evaluated on clean, noisy, and combined datasets, as well as with synthetic data augmentation. The results showed that training on both clean and noisy speech improves generalization, while augmentations such as SpecAugment, noise injection, and speed perturbation further enhance robustness--reducing the Word Error Rate by over 21%.

The findings highlight that integrating authentic and synthetic data is an effective approach for advancing ASR performance in low-resource languages. The BiGRU-based model achieved strong accuracy with lower computational overhead than BiLSTM counterparts, confirming its suitability for real-world and resource-constrained deployments.

Future work will explore multimodal ASR systems that integrate both audio and visual information, particularly lip movement cues, to enhance robustness against noise and improve intelligibility in challenging acoustic environments. This direction holds promise for developing inclusive and resilient speech technologies for underrepresented languages such as Amharic.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to Amplitude Ventures for supporting this research. In particular, we deeply appreciate the guidance and encouragement from Jakob Wredström, CEO of Amplitude Ventures, whose vision and leadership have been a source of inspiration throughout this work.

References

[1] L. Bahl, P. Brown, P. V. De Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 1986, pp. 49–52.

[2] H. A. Bourlard and N. Morgan, Connectionist Speech Recognition: A Hybrid Approach. Norwell, MA, USA: Kluwer Academic Publishers, 1993.

[3] T. Kebebew, "Speaker-dependent speech recognition for Afan Oromo using hybrid hidden Markov models and artificial neural network," M.S. thesis, Addis Ababa Univ., Addis Ababa, Ethiopia, 2010.

[4] A. Graves, Supervised Sequence Labelling with Recurrent Neural Networks, vol. 385, Studies in Computational Intelligence. Springer, 2012.

[5] M. B. Frew, "Audio-visual speech recognition using lip movement for Amharic language," Int. J. Eng. Res. Technol. (IJERT), vol. 8, no. 8, pp. 1–5, 2019.

[6] M. Yifiru, S. T. Abate, and W. Menzel, "Morpheme-based and factored language modeling for Amharic speech recognition," in Proc. Human Lang. Technol. Conf., 2009, pp. 1–5.

[7] H. O. Nasereddin and A. A. Hebash, "Classification techniques for automatic speech recognition (ASR) algorithms used with real-time speech translation," in Proc. 2017 Comput. Conf., London, U.K., 2018.

[8] G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition," IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, 2012.

[9] N. Jaitly and G. E. Hinton, "Learning a better representation of speech soundwaves using restricted Boltzmann machines," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2011, pp. 5884–5887.

[10] E. D. Emiru et al., "Improving Amharic speech recognition system using connectionist temporal classification with attention model and phoneme-based byte-pair encodings," Information, vol. 12, no. 2, p. 62, 2021.

[11] A. Gulati et al., "Conformer: Convolution-augmented Transformer for speech recognition," in Proc. Interspeech, 2020, pp. 5036–5040.

[12] D. Rekesh et al., "Fast Conformer with linearly scalable attention for efficient speech recognition," arXiv preprint arXiv:2305.05084, 2023. [Online]. Available: https://arxiv.org/abs/2305.05084

[13] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2014, pp. 1764–1772.

[14] H. Aldarmaki, A. Ullah, and N. Zaki, "Unsupervised automatic speech recognition: A review," arXiv preprint arXiv:2106.04897, 2021. [Online]. Available: https://arxiv.org/abs/2106.04897

[15] S. J. Basha, D. Veeraiah, B. V. Charan, W. S. Yeddu, and D. G. Babu, "Recent advancements in end-to-end speech recognition models: A survey," unpublished, 2023.

[16] J. Lee, J. Kang, and S. Watanabe, "Layer pruning on demand with intermediate CTC," in Proc. Interspeech 2021, 2021, pp. 3745–3749, doi: 10.21437/Interspeech.2021-1171.

[17] Y. A. Ejigu and T. T. Asfaw, "Enhancing Amharic speech recognition in noisy conditions through end-to-end deep learning," Preprints, 2024. https://doi.org/10.20944/preprints202402.0754.v1

[18] Y. A. Ejigu and T. T. Asfaw, "Large scale speech recognition for low resource language Amharic, an end-to-end approach," Preprints, 2024. https://doi.org/10.20944/preprints202402.0813.v1

[19] Y. Hono et al., "Integrating pre-trained speech and language models for end-to-end speech recognition," arXiv preprint arXiv:2312.03668, 2023. https://arxiv.org/abs/2312.03668

[20] Y. Bai et al., "Seed-ASR: Understanding diverse speech and contexts with LLM-based speech recognition," arXiv preprint arXiv:2407.04675, 2024. [Online]. Available: https://arxiv.org/abs/2407.04675

[21] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. Cubuk, and Q. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," Proc. Interspeech, 2019, pp. 2613–2617.

[22] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio Augmentation for Speech Recognition," in Proc. Interspeech 2015, 2015, pp. 3586-3589