Towards Speech Technology for Garo: A Low-Resource ASR System via Multilingual Transfer

Badal Nyalang * and Kathy Biginchi Ch Momin

Posted Date: 9 February 2026

doi: 10.20944/preprints202602.0686.v1

Keywords: Garo; automatic speech recognition; Whisper; low-resource languages; Northeast India; Tibeto-Burman; fine-tuning; speech technology

 doi:10.20944/preprints202602.0686.v1

Article

Towards Speech Technology for Garo: A Low-Resource ASR System via Multilingual Transfer

Badal Nyalang [1,*] and Kathy Biginchi Ch Momin [2]

[1] MWire Labs, Meghalaya, India

[2] Assam Don Bosco University, Assam, India

* Correspondence: nyalang@mwirelabs.com

Abstract

We present a fine-tuned Whisper model for automatic speech recognition (ASR) in Garo, a low-resource Tibeto-Burman language spoken in Northeast India. Using training samples from the Vaani dataset, we fine-tune Whisper-small and achieve a Word Error Rate (WER) of 9.74% and Character Error Rate (CER) of 3.82% on the test set, representing a 97.5% relative improvement over the zero-shot baseline. Our model produces perfect transcriptions for over 60% of test samples and achieves real-time inference speeds. We analyze error patterns including code-switching challenges and morphological complexities specific to Garo. The model is publicly released to support future research in low-resource speech recognition for Tibeto-Burman languages.

Keywords: Garo; automatic speech recognition; Whisper; low-resource languages; Northeast India; Tibeto-Burman; fine-tuning; speech technology

Introduction

Garo is a Tibeto-Burman language spoken by approximately 1.2 million people primarily in Meghalaya, India, and parts of Bangladesh. Despite its significant speaker population, Garo remains technologically underserved with limited digital resources and no publicly available speech recognition systems.

Automatic speech recognition has seen remarkable progress with transformer-based models like Whisper [1], which demonstrate strong multilingual capabilities. However, Whisper's training data does not include Garo, resulting in poor zero-shot performance.

In this work, we fine-tune Whisper-small on Garo speech data from the Vaani dataset [2] and evaluate its performance comprehensively. Our contributions include: (1) a publicly available ASR model for Garo language achieving under 10% WER; (2) comprehensive evaluation with detailed error analysis; (3) analysis of challenges specific to Garo phonology and morphology; and (4) open release of the trained model for community use.

Related Work

Low-Resource Speech Recognition

Low-resource ASR presents unique challenges due to limited training data, lack of standardized orthography, and scarce linguistic resources [3]. Transfer learning approaches have shown promise for adapting multilingual models to new languages with limited data [4,5]. Recent work demonstrates that pre-trained models can be effectively fine-tuned on datasets containing just thousands of utterances rather than the hundreds of thousands typically required for training from scratch [6,7].

Self-supervised learning methods like wav2vec [8] and wav2vec 2.0 [9] have reduced the amount of labeled data needed by pre-training on unlabeled speech. Fine-tuning multilingual models like

 doi:10.20944/preprints202602.0686.v1

Whisper offers a practical alternative, leveraging cross-lingual transfer from high-resource languages [10].

Speech Recognition for Indian Languages

Recent efforts have focused on developing ASR systems for low-resource Indian languages [11,12]. Work on Santali demonstrates the effectiveness of cross-lingual transfer learning with Whisper [13]. These studies highlight common challenges including limited training data, dialectal variation, and code-switching with dominant languages.

Garo Language Characteristics

Garo belongs to the Bodo-Garo branch of Tibeto-Burman languages. It has a relatively simple vowel system with six vowels but a complex consonant inventory including retroflex and palatal series. The language exhibits minimal tone but features distinctive vowel length and glottal stops. Garo is highly agglutinative, with extensive use of affixation, verbs can take multiple prefixes and suffixes indicating aspect, mood, and agreement. This agglutination creates long word forms that pose challenges for word-level error metrics. Garo uses a Latin-based script with additional diacritical marks, including the middle dot (·) for glottal stops and hyphens for certain morpheme boundaries.

Dataset

We use the Vaani-transcription-part dataset [2], a component of the larger Vaani project by ARTPARK-IISc. The dataset contains spontaneous, image-prompted speech collected across 165 districts in India, covering 109 languages. This collection methodology captures naturalistic speech patterns as speakers describe images shown to them.

The Garo subset is split into standard training, validation, and test partitions following an 80/8/10 ratio. Audio characteristics include a mean duration of approximately 4 seconds per sample at 16 kHz sampling rate. The relatively low type-token ratio in the dataset reflects Garo's agglutinative nature, where productive morphology creates diverse word forms from a smaller set of roots and affixes.

We apply the following preprocessing steps: removal of XML tags and bracketed corrections, conversion to lowercase, removal of punctuation except word-internal characters (·, -), and whitespace normalization. Audio is resampled to 16 kHz and converted to mel-spectrograms following Whisper's preprocessing pipeline.

Methodology

Whisper-small consists of 12 encoder and 12 decoder transformer layers with 244M total parameters and a 30-second context window. The encoder processes mel-spectrogram inputs while the decoder generates transcriptions autoregressively.

We fine-tune all model parameters using standard practices including AdamW optimizer, learning rate scheduling with warmup, mixed precision training (FP16), and gradient accumulation. Checkpoints are evaluated on the validation set, and the best performing checkpoint is selected based on validation loss convergence. Full training hyperparameters will be detailed in the extended version of this work.

Evaluation

We evaluate using two standard ASR metrics: Word Error Rate (WER) and Character Error Rate (CER). WER is computed as (S + D + I) / N where S = substitutions, D = deletions, I = insertions, and N = total words in reference. CER follows the same formulation at the character level. For fair comparison, we apply consistent normalization to both references and predictions including lowercase conversion, punctuation removal, and whitespace normalization.

 doi:10.20944/preprints202602.0686.v1

Results

Main Results

Table 1. Performance comparison with zero-shot baseline.

Model	WER (%)	CER (%)
Whisper-small (zero-shot)	382.7	203.5
GaroASR (fine-tuned)	**9.74**	**3.82**

Our fine-tuned model achieves a WER of 9.74% (97.5% relative improvement) and CER of 3.82% (98.1% relative improvement). The model produces perfect transcriptions (0% WER) for over 60% of test samples. The dramatic improvement over zero-shot performance demonstrates the effectiveness of fine-tuning for low-resource languages not represented in Whisper's training data.

Error Distribution

The zero median for both WER and CER indicates that most samples are transcribed perfectly, with errors concentrated in a subset of challenging cases. This bimodal distribution suggests the model has learned core Garo patterns effectively but struggles with specific phenomena.

Inference Speed

The model achieves a real-time factor of approximately 0.05× (20×faster than audio duration), enabling practical deployment for real-time applications including live transcription and voice interfaces.

Error Analysis

Error Patterns

Code-switching with English: Analysis of samples containing English loanwords reveals significant challenges, with a notable proportion exhibiting high error rates (WER > 30%). These errors often cascade, affecting surrounding Garo words as the model struggles to reconcile conflicting language patterns.

Annotation noise: A subset of samples contain bracketed corrections in the reference transcriptions, indicating transcriber uncertainty or dialectal variation. These corrections introduce ambiguity into evaluation, as the model may produce valid alternatives that differ from the corrected reference.

Compound word boundaries: Garo's agglutinative morphology creates compound forms with hyphens. The model occasionally segments these compounds incorrectly, reflecting tension between Garo's productive morphology and the word-tokenization assumptions of the underlying model.

CER vs WER Analysis

The CER/WER ratio of 0.39 indicates that errors are predominantly partial word mistakes rather than complete word substitutions or omissions. For agglutinative languages like Garo, a single morpheme error can invalidate an entire word from a WER perspective while preserving most characters. The low CER suggests the model captures Garo phonology and most morphological patterns effectively.

Discussion

The achieved WER of 9.74% represents strong performance for a low-resource language with no prior ASR systems. The 97.5% improvement over zero-shot performance validates fine-tuning as an effective strategy for low-resource ASR, consistent with findings from other low-resource languages

 doi:10.20944/preprints202602.0686.v1

[12,13]. The low CER/WER ratio (0.39) provides evidence that the model handles Garo's agglutinative morphology effectively.

The error rate on code-switched utterances highlights a significant limitation, likely reflecting limited English-Garo code-switching in training data, conflicting phonological and orthographic patterns between languages, and the model's bias toward producing Garo outputs. Code-switching is common in multilingual communities, making this a priority for future work.

This ASR system enables several practical applications for the Garo-speaking community including speech-to-text services, educational tools for language learning and literacy, documentation of oral traditions and cultural heritage, accessibility features for voice-controlled interfaces, and support for content creation in Garo language media. The real-time inference speed makes the model suitable for interactive applications.

Conclusions

We present a publicly available ASR system for Garo, achieving 9.74% WER and 3.82% CER through fine-tuning Whisper-small. The model produces perfect transcriptions for over 60% of test cases and operates at 20× real-time speed. Our error analysis reveals that code-switching with English and annotation noise present the primary challenges, while the model handles native Garo phonology and morphology effectively. By releasing this model publicly, we aim to support technological inclusion for the Garo-speaking community and contribute to the broader effort of developing language technologies for low-resource languages worldwide.

Limitations

This work has several limitations: high error rate on English loanwords limits applicability in code-switching contexts; training data may not capture full dialectal variation across Garo-speaking regions; annotation uncertainty affects a portion of test samples; only Whisper-small was evaluated, and larger model variants may achieve better performance; evaluation was conducted on a single test set from one collection methodology; and the model has not been tested on conversational or telephone speech.

Ethics Statement

This research aims to support linguistic diversity and technological inclusion for the Garo-speaking community. The Vaani dataset was collected with appropriate consent and ethical approval from participants. We release this model openly to benefit speakers of Garo and researchers working on low-resource languages. We acknowledge that ASR technology can be used for both beneficial and harmful purposes, including surveillance. We encourage responsible deployment that respects speaker privacy, cultural context, and community needs.

Funding: This research was funded by MWire Labs.

Data Availability Statement: The Vaani dataset used in this study is publicly available at https://huggingface.co/datasets/ARTPARK-IISc/Vaani-transcription-part. The trained model is publicly released for community use.

Acknowledgments: We thank ARTPARK-IISc for creating and releasing the Vaani dataset, which made this work possible. We also acknowledge the Garo-speaking participants who contributed their speech to the dataset.

Conflicts of Interest: The authors declare no conflict of interest.

 doi:10.20944/preprints202602.0686.v1

References

1. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. In Proceedings of the International Conference on Machine Learning, 2023; pp. 28492–28518.
2. ARTPARK-IISc. Vaani: Capturing the true diversity of India's spoken languages. 2024. Available online: https://huggingface.co/datasets/ARTPARK-IISc/Vaani-transcription-part (accessed on 22 January 2025).
3. Besacier, L.; Barnard, E.; Karpov, A.; Schultz, T. Automatic speech recognition for under-resourced languages: A survey. Speech Commun. 2014, 56, 85–100.
4. Conneau, A.; Baevski, A.; Collobert, R.; Mohamed, A.; Auli, M. Unsupervised cross-lingual representation learning for speech recognition. In Proceedings of Interspeech, 2020; pp. 2426–2430.
5. Gisslen, N.R.; Hedlund, E.F.; Brown, S.; Bird, S. Breaking the transcription bottleneck: Fine-tuning ASR models for extremely low-resource fieldwork languages. In Proceedings of the 6th Workshop on the Use of Computational Methods in the Study of Endangered Languages, 2025.
6. Pratap, V.; Tjandra, A.; Shi, B.; Tomasello, P.; Babu, A.; Kundu, S.; et al. Scaling speech technology to 1,000+ languages. arXiv 2023, arXiv:2305.13516.
7. Prajapati, A.; Kumar, K.; Jyothi, P. 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), 2024.
8. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. wav2vec: Unsupervised pre-training for speech recognition. In Proceedings of Interspeech, 2019; pp. 3465–3469.
9. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems, 2020; Volume 33, pp. 12449–12460.
10. Zhang, Y.; Park, D.S.; Han, W.; Qin, J.; Gulati, A.; Shor, J.; et al. Google USM: Scaling automatic speech recognition beyond 100 languages. arXiv 2023, arXiv:2303.01037.
11. Prabhavalkar, R.; Rao, K.; Sainath, T.N.; Li, B.; Johnson, L.; Jaitly, N. Multilingual speech recognition for Indian languages. In Intelligent Computing: Proceedings of the 2022 Computing Conference; Springer, 2022; pp. 528–542.
12. Sinha, S.; Saabith, A.L.S.; Jyothi, P. Model adaptation for ASR in low-resource Indian languages. arXiv 2023, arXiv:2307.07948.
13. Muralidaran, D.; Singh, Y.K.; Kumar, R. Whispering in Ol Chiki: Cross-lingual transfer learning for Santali speech recognition. In Findings of the Association for Computational Linguistics: IJCNLP-AACL, 2025.