

NE-MultiSpeech: A Multilingual Speech Corpus and ASR Benchmark for Northeast Indian Languages

Badal Nyalang

MWire Labs, Meghalaya, India
nyalang@mwirelabs.com

Angana Borah

University of Michigan, Ann Arbor, USA
anganab@umich.edu

Abstract

NE-MultiSpeech is a large-scale transcribed multilingual speech corpus covering eight indigenous and regional languages of Northeast India: Khasi, Garo, Mizo, Nagamese, Kokborok, Assamese, Wanchó, and Chakma. The corpus combines 124 hours of newly collected, speaker-compensated proprietary recordings across five languages with curated subsets from the Vaani dataset (Pulikodan et al., 2026), totalling approximately 267 hours of transcribed speech across three language families. We fine-tune Whisper-medium on NE-MultiSpeech and establish a systematic automatic speech recognition (ASR) benchmark across all eight languages, achieving an overall word error rate (WER) of 35.53% over 8,099 test utterances jointly, outperforming both Whisper-large-v3 zero-shot (118–194% WER) and MMS-1B zero-shot (42–116% WER) by a large margin. We additionally introduce NE-SpeechEmbed, a dual-encoder speech-text retrieval model trained on 73,476 speech-text pairs via InfoNCE contrastive learning, establishing a retrieval baseline across these languages, consistently outperforming CLAP zero-shot on all eight languages at full test-pool scale (approximately 1,000 candidates per language). Probing the fine-tuned encoder with t-SNE reveals language-family-aligned clustering tendencies without explicit family supervision. All models are released under CC-BY-4.0; data splits and utterance IDs are released under CC-BY-4.0, with proprietary audio available under a research data access agreement.

1 Introduction

Northeast India is home to one of the world’s highest concentrations of linguistic diversity, with over 220 languages from four major families spanning eight states. Austroasiatic languages such as Khasi are spoken by over one million people in Meghalaya; Tibeto-Burman languages including Garo, Mizo, Kokborok, Wanchó, and

Chakma span Meghalaya, Mizoram, Tripura, and Arunachal Pradesh; Indo-Aryan languages including Assamese and the creole Nagamese serve as lingua francas across the region. Despite their cultural significance, nearly all of these languages remain entirely absent from modern speech technology.

The consequences are concrete. Voice interfaces, government digital services, healthcare delivery, and educational technology built on ASR remain inaccessible to large portions of the Northeast Indian population. The primary bottleneck is data scarcity: no unified, speaker-disjoint transcribed speech corpus spanning all eight languages in our study exists.

Existing multilingual resources are insufficient. FLEURS (Conneau et al., 2023) and MLS (Pratap et al., 2020) include Assamese at best. Whisper (Radford et al., 2023) covers none of our target languages in its 99-language set except for Assamese. MMS (Pratap et al., 2024), despite nominally supporting over 1,100 languages, lacks adapters for four of the eight languages we study. Large audio-language models such as SALMONN (Tang et al., 2024) and Qwen2-Audio (Chu et al., 2024) offer strong general-purpose speech understanding but do not address the fundamental data gap for languages absent from their training corpora.

We make the following contributions:

- **NE-MultiSpeech:** a 267-hour transcribed multilingual speech corpus across eight Northeast Indian languages from three language families, with speaker-disjoint train/val/test splits for all proprietary languages (Figure 1).
- **NE-ASR:** a fine-tuned Whisper-medium model with a principled proxy language token strategy for unsupported languages, empirically ablated on Khasi across five candidate proxies, achieving overall WER 35.53% and

outperforming Whisper-large-v3 zero-shot on every language (overall 162.69% $\rightarrow$ 35.53% WER; 4.6 $\times$).

- **NE-SpeechEmbed:** a speech-text retrieval model for Northeast Indian languages, benchmarked against CLAP zero-shot across all eight languages and three language families.
- **NE-MMS:** adapter-only fine-tuning of MMS-1B across all eight languages, providing a direct comparison of full fine-tuning vs. adapter-only fine-tuning for low-resource ASR.
- Encoder probing results suggesting representations partially aligned with language-family structure without explicit supervision (Figure 3).
- **NortheastBench-Speech:** a standardized 8,099-utterance evaluation suite across 8 languages and 3 families for ASR and speech-text retrieval, with full public release of models, data splits, and utterance IDs.

2 Related Work

Multilingual ASR. Whisper (Radford et al., 2023) trains on 680,000 hours of weakly supervised web audio spanning 99 languages. MMS (Pratap et al., 2024) extends ASR coverage to 1,100+ languages via language-specific adapters on a shared wav2vec 2.0 backbone. Wav2Vec 2.0 XLSR-53 (Conneau et al., 2021) demonstrates cross-lingual transfer to low-resource languages from self-supervised pretraining. SeamlessM4T (Seamless Communication et al., 2023) provides unified multilingual speech and text translation across nearly 100 languages. None of these systems provide usable coverage of the Northeast Indian languages studied here.

Indic Speech Resources. Project Vaani (Pulikodan et al., 2026) collects spontaneous image-prompted speech from 165 districts across India, covering 105 languages including several from Northeast India. IndicVoices (Javed et al., 2024) and IndicVoices-R (Sankar et al., 2024) provide large spontaneous Indic speech datasets. Vistaar (Bhogale et al., 2023) provides diverse ASR benchmarks for Indic languages. Vividh-ASR (Juvekar et al., 2026) introduces a complexity-tiered Whisper fine-tuning benchmark, identifying studio-bias as a key failure mode; we observe a related

pattern in our proprietary vs. Vaani ablation results. No prior work benchmarks ASR systematically across the eight languages we study.

Contrastive Speech-Text Learning.

CLAP (Elizalde et al., 2023) aligns audio and text via dual encoders and InfoNCE loss (van den Oord et al., 2018). Sankaran et al. (2026) demonstrate few-shot contrastive adaptation of CLAP for low-resource Indic audio using projection-only fine-tuning, the same strategy we adopt for NE-SpeechEmbed. SALMONN (Tang et al., 2024) and Qwen2-Audio (Chu et al., 2024) represent recent large audio-language models; NE-SpeechEmbed targets a complementary low-resource, multilingual, transcript-level retrieval setting where such models lack training coverage.

Whisper Adaptation. Pillai et al. (2026) propose multistage fine-tuning schedules for low-resource languages. Timmel et al. (2025) analyze practical Whisper adaptation including language token selection, directly motivating our proxy token ablation.

Multilingual Speech Benchmarks.

MuAViC (Anwar et al., 2023) provides a multilingual audio-visual corpus for robust speech recognition; NortheastBench-Speech addresses a complementary gap for extremely low-resource languages with no prior benchmark coverage.

3 NE-MultiSpeech Dataset

3.1 Proprietary Corpus

We collected a new speech corpus for five Northeast Indian languages (Table 1) through a structured community data collection program. Native speakers were recruited locally and compensated at Rs. 400 per recorded hour, equivalent to approximately 4.6 $\times$ the mean daily wage across Northeast Indian states (mean Rs. 351/day across Meghalaya, Assam, Mizoram, Tripura, Nagaland, Manipur, and Arunachal Pradesh; source: Data Commons), well above prevailing rates for data collection work in the region. Speakers provided written informed consent in their native language prior to recording, administered digitally via Google Forms; the English translation of the consent instrument is in Appendix E. Recordings were made using consumer smartphones in naturalistic indoor environments. Transcripts, splits, and utterance IDs are released under CC-BY-4.0; audio is available via a free, one-click access agreement (Section 7). Train, validation, and test splits are strictly speaker-disjoint.

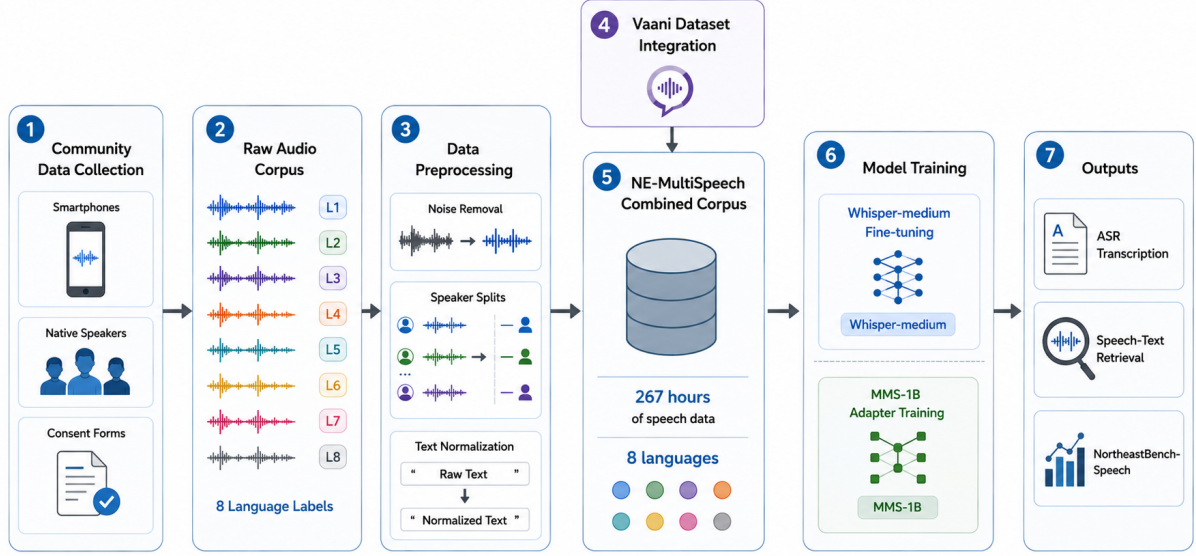


Figure 1: NE-MultiSpeech pipeline: from community data collection and Vaani integration through preprocessing to model training and outputs.

Lang.	Hrs	Utt.	Spk.	Mean (s)
Khasi	67.70	23,766	49	10.25
Nagamese	17.02	8,902	6	6.88
Mizo	14.21	8,027	6	6.37
Garó	13.05	6,945	6	6.77
Kokborok	12.08	3,572	12	12.17
Total	124.06	51,212	79	—

Table 1: Proprietary NE-Speech corpus. All splits are speaker-disjoint.

Khasi has extreme duration outliers (maximum 1,804s, mean 10.25s); a duration filter removing utterances below 1s and above 30s was applied before training.

3.2 Vaani Subsets

We extract and clean language-specific subsets from Vaani (Pulikodan et al., 2026) for seven languages. Vaani collects spontaneous image-prompted speech from 165 districts across India, with multi-stage quality control including SNR filtering and native-speaker transcription validation. Our preprocessing removes noise markers, correction annotations, and bracket markers. Validation and test splits are capped at 1,000 utterances per language; Kokborok is limited by source availability (279 test utterances). Full split counts are in Appendix B. Audio is not redistributed; we release utterance IDs under CC-BY-4.0, with audio accessed via the original Vaani repository. Vaani’s collec-

tion pipeline additionally enforces speaker-level deduplication (Pulikodan et al., 2026); we follow Vaani’s provided split assignments directly and do not reshuffle, preserving whatever speaker-level separation is enforced by the source dataset. For the four languages appearing in both Vaani and our proprietary corpus (Garó, Mizo, Nagamese, Kokborok), proprietary and Vaani speakers are entirely disjoint by construction, as they were recruited independently.

3.3 Combined Corpus and Language Overview

NE-MultiSpeech combines both sources, yielding 267 hours across eight languages from three language families (Table 2). Garó, Mizo, Nagamese, and Kokborok appear in both sources; training data is pooled. The dataset spans Austroasiatic, Tibeto-Burman, and Indo-Aryan families, with both Roman and native script orthographies.

4 NE-ASR: Automatic Speech Recognition

4.1 Proxy Language Token Strategy

Whisper’s multilingual decoder requires a language token from a fixed set of 99 supported languages. None of our eight target languages appear in this set except Assamese. Training without a forced language token causes unstable WER across training steps due to Whisper’s internal language detector.

Lang.	Family	Script	Src.	Spk.
Khasi	Austroasiatic	Roman	Prop.	49
Garó	Tibeto-Burman	Roman	Both	6
Mizo	Tibeto-Burman	Roman	Both	6
Nagamese	Indo-Aryan (Cr.)	Roman	Both	6
Kokborok	Tibeto-Burman	Roman	Both	12
Assamese	Indo-Aryan	Assamese	Vaani	–
Wancho	Tibeto-Burman	Roman	Vaani	–
Chakma	Tibeto-Burman	Roman	Vaani	–

Table 2: Language overview. Nagamese is an Indo-Aryan creole (Cr.). Speaker counts for proprietary languages only.

Proxy	Stable	WER	Notes
None (auto)	No	18.16	Oscillates
Welsh cy	Yes	14.94	Best
Khmer km	Yes	24.29	Austroasiatic
Vietnamese vi	Yes	106.77	Catastrophic
Custom < kh >	No	–	Decoder failure

Table 3: Proxy token ablation on Khasi. Custom token initialized from the mean of Khmer and Vietnamese embeddings.

We conducted a systematic ablation on Khasi, evaluating five proxy strategies (Table 3). Welsh (cy) achieved the best training stability and lowest WER. Khmer (km), despite being Austroasiatic like Khasi, performed worse, suggesting that linguistic family proximity does not reliably predict proxy token quality in Whisper. Vietnamese (vi) caused catastrophic hallucination. A custom <|kh|> token initialized from the mean of Khmer and Vietnamese embeddings failed entirely: the decoder was unable to acquire a new language token during fine-tuning without full architecture retraining. This negative result is practically important: proxy selection cannot be resolved by language family heuristics alone.

Based on these results we derive a unified proxy strategy for all eight languages (Table 5). Assamese uses its native Whisper token. Tibeto-Burman languages use Burmese (my). Khasi, Nagamese, and Wancho use Welsh (cy) for stable Roman-script generation. The same token is forced consistently during both training and inference; inconsistency between phases was a primary failure mode in early experiments.

For languages beyond Khasi, proxy selection followed two criteria derived from the Khasi ablation: training stability (consistent WER convergence across steps) and script compatibility (Roman-script proxies for Roman-script targets).

Lang.	None	my	cy	vi	Best
Garó	51.78	46.08	50.59	48.05	my
Mizo	33.32	31.92	32.74	33.90	my
Kokborok	34.56	31.71	33.91	34.71	my

Table 4: Proxy ablation on three Tibeto-Burman languages (none / Burmese my / Welsh cy / Vietnamese vi). Family-matched Burmese is best in all three.

Language	Proxy	Rationale
Khasi	Welsh cy	Roman, ablated best
Garó	Burmese my	Tibeto-Burman
Mizo	Burmese my	Tibeto-Burman
Nagamese	Welsh cy	Roman script
Kokborok	Burmese my	Tibeto-Burman
Assamese	Assamese as	Native token
Wancho	Welsh cy	Roman transliteration
Chakma	Burmese my	Tibeto-Burman

Table 5: Final proxy language token assignments.

Burmese (my) was selected for Tibeto-Burman languages as the best-represented family member in Whisper’s 99-language set. Welsh (cy) was applied to all Roman-script languages without a native token (Nagamese, Wancho), since the Khasi ablation established its stability advantage over other Roman-script candidates. To validate this generalization directly, we ran per-language proxy ablations for three Tibeto-Burman languages (Garó, Mizo, Kokborok), each tested with four proxy tokens (none, Burmese, Welsh, Vietnamese), holding data and hyperparameters fixed within each language (Table 4). All three confirm the family-matched Burmese proxy as best, consistent with our Table 5 assignment. This differs from Khasi (Austroasiatic), where the non-family Welsh proxy won, indicating proxy quality is language- and family-dependent rather than governed by a single universal rule.

4.2 Model and Training

We fine-tune Whisper-medium (Radford et al., 2023) (307M parameters) jointly on all eight languages. Training uses learning rate 1e-5 with linear warmup over 500 steps, gradient accumulation of 2 (effective batch size 32), fp16 precision, and AdamW for 8,000 steps on a single NVIDIA A100 80GB. Training pairs are capped at 10,000 per language; Kokborok is underrepresented at 6,083 combined pairs due to limited Vaani availability. Full training and decoding details are in Appendix A.

4.3 Evaluation and Baselines

We evaluate on held-out test sets: 1,000 utterances for Khasi, 1,099 for Kokborok, and 1,000 from the Vaani test split for all other languages (8,099 total). All conditions – NE-ASR fine-tuned, Whisper-large-v3 zero-shot, and MMS-1B zero-shot – use identical greedy decoding settings: beam size 1, temperature 0.0, and task=transcribe, ensuring a fair comparison across baselines. WER is the primary metric; overall WER is computed over all languages jointly. CER is reported on a 100-sample subset per language.

We compare against three baselines. **Whisper-large-v3** uses greedy decoding, temperature 0.0, and the same proxy token strategy as NE-ASR. **MMS-1B zero-shot** uses native adapters where available and typologically motivated proxies otherwise. **MMS-1B fine-tuned** trains only the MMS-1B adapter layers on NE-MultiSpeech data (50–180K parameters per language, backbone frozen), providing a direct comparison against full fine-tuning of Whisper-medium (307M parameters updated).

WER exceeds 100% when a model hallucinates output in the proxy language, producing more words than the reference. Manual inspection of 20 samples per language confirms all $\text{WER} > 100\%$ cases correspond to proxy language hallucination.

4.4 Analysis

Data fusion. The combined model outperforms either source alone for Garo, Mizo, Kokborok, Assamese, and Chakma. The prop-only condition produces WER above 100% for Kokborok (118.83%); adding Vaani drops this to 44.79%. Khasi achieves its best WER under prop-only (14.94%), reflecting the size and quality of its 23,766-utterance corpus. For Assamese and Chakma, which have no proprietary data, the combined (jointly multilingual) checkpoint still outperforms the Vaani-only single-language WER (Assamese: 20.98% vs. 30.02%; Chakma: 54.25% vs. 57.25%), indicating positive cross-lingual transfer from joint training on the other six languages.

Zero-shot gap and fine-tuning gains. Fine-tuning Whisper-medium, a $4\times$ smaller model than Whisper-large-v3, reduces WER on every language (overall 162.69% $\rightarrow$ 35.53%; $4.6\times$), with the largest gain on Garo ($20.4\times$) and the smallest on Wancho ($2.4\times$). MMS-1B outperforms Whisper-large-v3 zero-shot for the four languages with na-

tive adapters but still substantially underperforms NE-ASR.

Full fine-tuning vs. adapter-only. NE-ASR outperforms MMS-1B fine-tuned on every language. The gap is largest for Kokborok and Chakma, where adapter fine-tuning does not reduce WER below 100%. Even for languages with strong native MMS adapters, full fine-tuning wins: Assamese (39.30% vs. 20.98%), Garo (40.24% vs. 9.31%). These results suggest that for extremely low-resource languages, full fine-tuning of a well-pretrained model tends to be a more effective strategy than adapter-only fine-tuning of a larger model, though the comparison is not fully controlled for model size.

MMS adapter script mismatch. The MMS-1B native Garo adapter (grt) produces Bengali-script output despite Latin-script input, because the adapter was trained on the historical Bengali-Assamese orthography of Garo. Table 7 shows representative outputs: the transcriptions are phonetically plausible but yield 100%+ CER against Latin references. Substituting the Mizo (miz) adapter as a Latin-script Tibeto-Burman proxy reduces WER to 40.24%. The full Bengali-script outputs are shown in Appendix C. This illustrates a practical gap between claimed language coverage and deployment readiness that may affect other languages in the MMS inventory.

Training data size vs. WER. Figure 2 plots combined training hours against WER. Garo is a positive outlier (46 hours, 9.31% WER) driven by the large Vaani corpus (33,480 utterances). Wancho has the fewest hours (8.42h) and highest WER (68.37%). Assamese departs from the trend with moderate hours but low WER (20.98%), benefiting from its native Whisper token. Kokborok is a negative outlier despite moderate hours, reflecting its Vaani subset of only 2,511 utterances.

CER analysis. CER is substantially lower than WER across all languages, indicating phonetically plausible output even when word boundaries are incorrect. Chakma and Kokborok show the largest WER-CER gap, consistent with agglutinative morphology. Assamese CER (17.91%) closely tracks its WER (20.98%), suggesting script-level difficulty beyond word segmentation.

Cross-family transfer. Four Tibeto-Burman languages (Garo, Mizo, Kokborok, Chakma) share a

Language	NE-ASR (Whisper-medium FT)			Zero-shot		MMS-1B FT		CER [†]
	Vaani	Prop	Comb.	Wh-lv3	MMS-1B	WER	CER	
Khasi	—	14.94	16.89	136.92	100.00*	54.19	15.00	7.05
Garó	11.51	63.43	9.31	189.91	116.49	40.24 [‡]	8.81 [‡]	3.80
Mizo	30.51	69.10	23.85	142.97	84.93	61.71	25.11	14.92
Nagamese	47.53	86.43	49.13	172.69	77.84	72.29	31.58	26.82
Kokborok	66.77	118.83	44.79	183.56	99.97*	100.00	46.57	26.94
Assamese	30.02	—	20.98	118.33	42.30	39.30	16.54	17.91
Chakma	57.25	—	54.25	193.55	100.00*	100.00	38.96	17.82
Wancho	63.76	—	68.37	163.60	100.00*	83.67	30.81	25.99
Overall	41.88	—	35.53	162.69	90.18	—	—	—

Table 6: WER (%) and CER (%) results. Bold: best NE-ASR condition per language. Overall WER over all 8,099 test utterances jointly. [†]CER on 100-sample subset (NE-ASR combined condition). *No native MMS-1B adapter; typologically motivated proxy used. [‡]Garó MMS-1B uses Mizo (miz) proxy due to grt adapter script mismatch (Section 4.4).

Reference (Latin Garó)	Transliteration of MMS output	Commentary
ia noksao kosako see tape donaniko nikna gita manenga	iya noksao khosokho se tappe do- nanokho nikna gitta ma nningya	Phonetically faithful; wrong script
ian gari tarichakram biap ba nok ba dokan onga	yan gadi thari chakram bi a ba nok ba dokkhan ha	Partial word splits; correct sounds
banga rasinkoba iano nikna gita mana	bang a rasinkoba iyano nikna gitta ma nna	Near-perfect phoneme mapping
ia noksao anga mechik saksani chrokaniko nika	iya noksao anga mechik saksani chrokanokho nikka	Structure preserved; script wrong
iano bolrangkoba nikatna gita ma- nenga	iyano bolrangkoba nikkhatna gitta ma nnenga	Consistent Bengali rendering

Table 7: MMS-1B grt adapter outputs transcribed into Latin transliteration. The model produces phonetically plausible Garó sounds but in Bengali script, yielding 100%+ CER against Latin references. Full Bengali-script outputs are in Appendix C.

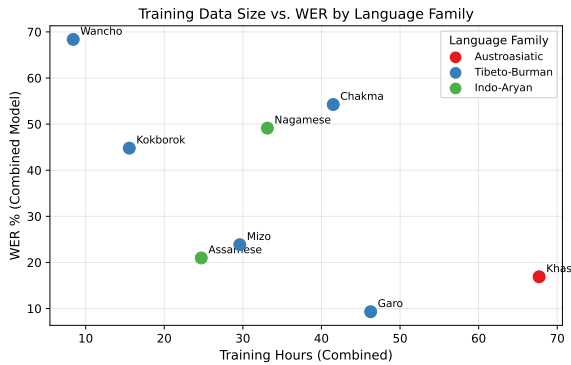


Figure 2: Combined training hours vs. WER (%). Garó is a positive outlier driven by large clean Vaani data. Kokborok is a negative outlier due to limited Vaani utterance count.

Burmese proxy token; Wancho uses Welsh. All eight languages are trained jointly. Garó achieves the lowest WER (9.31%). Wancho and Chakma lag despite sharing the family, reflecting smaller training sets with no proprietary supplement.

5 NE-SpeechEmbed: Speech-Text Retrieval

5.1 Model Architecture

NE-SpeechEmbed is a dual-encoder contrastive model for speech-to-text retrieval. The speech encoder is the encoder stack of our NE-ASR checkpoint (Whisper-medium, 1024-dim), projected via a learned linear layer to a 768-dim shared space. The text encoder is xlm-roberta-base (Conneau et al., 2020) (CLS token, 768-dim), projected to the same space. Both embeddings are L2-normalized. All encoder parameters are frozen; only the two projection layers (1,377,793 parameters) and a learned log-temperature scalar are trained via symmetric InfoNCE loss (van den Oord et al., 2018). We use AdamW (lr=2e-5, batch size=64, gradient clip=1.0) for 10 epochs on a single NVIDIA A100 80GB.

5.2 Training Data and Evaluation

We construct 73,476 speech-text pairs across eight languages, capped at 10,000 per language; Kok-

borok is underrepresented at 4,616 pairs. Full pair counts are in Appendix B. We evaluate speech-to-text retrieval using Recall at K ($R@K$) on the full test pool per language (approximately 1,000 candidates, 1,099 for Kokborok), comparing against CLAP (Elizalde et al., 2023) zero-shot using audio padded or truncated to a fixed 10-second window.

5.3 Results and Analysis

NE-SpeechEmbed outperforms CLAP zero-shot on every language and metric at full-pool scale (1.4M trainable parameters, $\sim 10,000$ pairs per language), with CLAP falling to or below random on several languages (Mizo, Chakma $R@1$). Kokborok achieves the highest $R@1$ (1.18%) despite being the most data-limited language; Mizo the lowest (0.2%). The margin over CLAP holds at full-pool scale and is not an artifact of the smaller 100-sample pool used in preliminary evaluation. Results hold across all three language families with no cross-family collapse, suggesting the shared embedding space may generalise across typologically diverse families, though further work is needed to confirm this at scale.

5.4 Encoder Probing: Language Family Structure

We extract mean-pooled encoder embeddings from 50 test utterances per language (400 total) and apply t-SNE dimensionality reduction (van der Maaten and Hinton, 2008). Figure 3 shows the resulting visualization. Each language forms a relatively tight cluster. More notably, the three language families tend to occupy distinct spatial regions without explicit family-level supervision during training, suggesting that the fine-tuned encoder may develop representations partially aligned with language-family structure. Khasi (Austroasiatic) appears largely isolated; the two Indo-Aryan languages (Nagamese, Assamese) cluster in a shared region; the five Tibeto-Burman languages occupy the center with some within-family sub-clustering. We note that t-SNE visualizations are sensitive to hyperparameter choices and should be interpreted cautiously; nonetheless, the broad family-level separation is consistent across multiple random seeds. To quantify this beyond the 2D projection, we compute the silhouette score on the original 1024-dim encoder embeddings: 0.11 for family labels and 0.44 for language labels (cosine distance, 400 embeddings across 8 languages), confirming real but modest family-level structure alongside strong

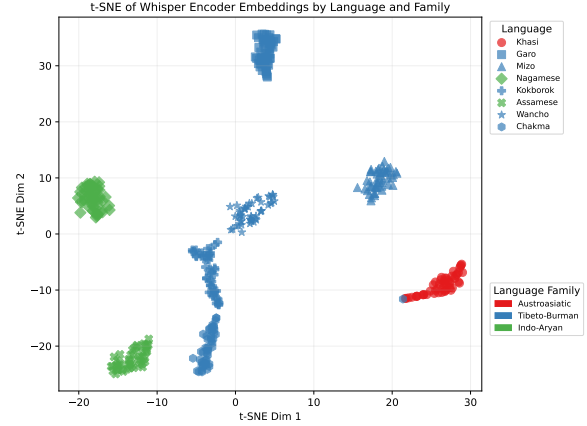


Figure 3: t-SNE of mean-pooled Whisper encoder embeddings (50 utterances per language, 400 total). Language families tend to occupy distinct spatial regions without explicit supervision, though t-SNE results should be interpreted with care.

language-level separation.

6 Ethical Considerations

All proprietary data was collected with the full informed consent of speakers, provided in their native language prior to recording via Google Forms. Speakers were compensated at Rs. 400/hour (approximately $4.6\times$ the mean NE state daily wage) and explicitly notified of intended uses. Transcripts, splits, and utterance IDs are released under CC-BY-4.0 with speaker identity anonymised; audio access is described in Section 7. The consent instrument is reproduced in Appendix E.

The languages in this study are spoken by indigenous and minority communities historically marginalised from technology development. We view this corpus as a step toward reducing that gap and encourage future work to involve community stakeholders in annotation and downstream application design. ASR technology built on this data could be misused for surveillance; we note this as a consideration for downstream deployers.

7 Data and Model Release

We release **NE-ASR** (fine-tuned Whisper-medium) at <https://huggingface.co/MWirelabs/ne-asr> and **NE-SpeechEmbed** (dual-encoder retrieval model) at <https://huggingface.co/MWirelabs/ne-speechembed>, both under CC-BY-4.0. **NortheastBench-Speech** (8,099-utterance benchmark) is released at <https://huggingface.co/datasets/MWirelabs/northeastbench-speech>. The

Language	NE-SpeechEmbed			CLAP (zero-shot)			Random
	R@1	R@5	R@10	R@1	R@5	R@10	R@1
Khasi	0.8	3.1	5.6	0.3	1.3	1.8	0.1
Garo	0.3	1.9	3.6	0.2	0.5	1.4	0.1
Mizo	0.2	1.1	2.7	0.0	0.7	1.3	0.1
Nagamese	0.5	2.2	4.8	0.2	0.5	0.9	0.1
Kokborok	1.18	4.00	7.92	0.09	0.91	1.27	0.09
Assamese	0.3	2.3	3.9	0.1	0.6	1.0	0.1
Wancho	0.5	3.6	6.5	0.1	0.4	0.8	0.1
Chakma	1.0	3.8	7.1	0.0	0.1	0.7	0.1

Table 8: Speech-text retrieval, full test pool per language (approximately 1,000 candidates, 1,099 for Kokborok). NE-SpeechEmbed outperforms CLAP zero-shot on all languages and metrics; CLAP performs at or below random for several languages.

proprietary corpus (transcripts, splits, and utterance IDs) is released under CC-BY-4.0 at <https://huggingface.co/datasets/MWirelabs/northeast-india-voices>. Audio access requires only a one-click acceptance of a free, automated usage agreement at the same URL, used for usage tracking rather than gatekeeping; no manual review or approval step is involved. Vaani utterance IDs are released under CC-BY-4.0; audio is accessed via the original Vaani repository.

8 Conclusion

NE-MultiSpeech provides a large-scale transcribed multilingual speech corpus for Northeast Indian languages, covering eight languages from three language families across 267 hours. Fine-tuning Whisper-medium yields a systematic ASR benchmark for the region, outperforming Whisper-large-v3 zero-shot on every language (overall 162.69% $\rightarrow$ 35.53% WER; $4.6\times$). NE-SpeechEmbed establishes a speech-text retrieval baseline, consistently outperforming CLAP zero-shot. A comparison of fine-tuning strategies suggests that full fine-tuning of Whisper-medium tends to outperform adapter-only fine-tuning of MMS-1B across all languages, particularly for those without native adapter support. We additionally document an MMS adapter script mismatch for Garo, where the native adapter targets Bengali script despite modern Latin-script usage, as a practical example of the gap between claimed multilingual coverage and deployment readiness. Encoder probing reveals representations partially aligned with language-family structure, consistent with the hypothesis that joint multilingual fine-tuning may preserve family-level phonetic organization. As large audio-language models such as SALMONN (Tang et al., 2024) and Qwen2-Audio (Chu et al., 2024) continue to

advance general speech understanding, the data gap for extremely low-resource language families remains a binding constraint; NE-MultiSpeech provides a foundation to begin addressing it for North-east India.

Limitations

WER remains high for Nagamese (49.13%), Wancho (68.37%), and Chakma (54.25%), reflecting limited training data. NE-SpeechEmbed should be understood as a baseline given its small training data relative to English CLAP systems. Speaker diversity is limited for some proprietary languages (6 speakers for Garo, Mizo, Nagamese). Temperature-based language sampling, cross-language hard negatives, and monolingual fine-tuning ablations are left as future directions. The fine-tuning strategy comparison (full vs. adapter-only) is not fully controlled for model capacity; future work should isolate this variable.

Acknowledgments

We thank the native speaker contributors to the proprietary corpus for their time and generosity, and the Vaani team at ART PARK-IISc for open data release. We also thank the reviewers and area chair for their detailed, constructive feedback, which substantially improved this paper.

References

- Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, and Changhan Wang. 2023. [MuAViC: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation](#). In *Proceedings of Interspeech 2023*, pages 4064–4068.
- Kaushal Bhogale, Sai Sundaresan, Abhigyan Raman, Tahir Javed, Mitesh M. Khapra, and Pratyush Kumar.

2023. [Vistaar: Diverse benchmarks and training sets for indian language ASR](#). In *Proceedings of Interspeech 2023*, pages 4384–4388.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. [Qwen2-audio technical report](#). *arXiv preprint arXiv:2407.10759*.
- Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. [Unsupervised cross-lingual representation learning for speech recognition](#). In *Proceedings of Interspeech 2021*, pages 2426–2430.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.
- Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. [FLEURS: Few-shot learning evaluation of universal representations of speech](#). In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.
- Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. [CLAP: Learning audio concepts from natural language supervision](#). In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Tahir Javed, Janki Nawale, Eldho George, Sakshi Joshi, Kaushal Bhogale, Deovrat Mehendale, Ishvinder Sethi, Aparna Ananthanarayanan, Hafsa Faquih, Pratiti Palit, Sneha Ravishankar, Saranya Sukumaran, Tripura Panchagnula, Sunjay Murali, Kunal Gandhi, Ambujavalli R, Manickam M, C Vijayanthi, Krishnan Karunganni, and 2 others. 2024. [IndicVoices: Towards building an inclusive multilingual speech dataset for Indian languages](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10740–10782.
- Kush Juvekar, Kavya Manohar, Aditya Menon, Arghya Bhattacharya, and Kumarmanas Nethil. 2026. [Vividh-ASR: A complexity-tiered benchmark and optimization dynamics for robust indic speech recognition](#). *arXiv preprint arXiv:2605.13087*.
- Leena G Pillai, Kavya Manohar, Basil Raju, and Elizabeth Sherly. 2026. [Multistage fine-tuning strategies for automatic speech recognition in low-resource languages](#). *ACM Transactions on Asian and Low-Resource Language Information Processing*.
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. [Scaling speech technology to 1,000+ languages](#). *Journal of Machine Learning Research*, 25(97):1–52.
- Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. [MLS: A large-scale multilingual dataset for speech research](#). In *Proceedings of Interspeech 2020*, pages 2757–2761.
- Sujith Pulikodan, Abhayjeet Singh, Agneedh Basu, Nihar Desai, Pavan Kumar J, Pranav D Bhat, Raghu Dharmaraju, Ritika Gupta, Sathvik Udapa, Saurabh Kumar, Sumit Sharma, Visruth Sanka, Dinesh Tewari, Harsh Dhand, Amrita Kamat, Sukhwinder Singh, Shikhar Vashishth, Partha Talukdar, Raj Acharya, and Prasanta Kumar Ghosh. 2026. [VAANI: Capturing the language landscape for an inclusive digital India](#). *Preprint*, arXiv:2603.28714.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. [Robust speech recognition via large-scale weak supervision](#). In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518.
- Ashwin Sankar, Srija Anand, Praveen Srinivasa Varadhan, Sherry Thomas, Mehak Singal, Shridhar Kumar, Deovrat Mehendale, Aditi Krishana, Giri Raju, and Mitesh Khapra. 2024. [IndicVoices-R: Unlocking a massive multilingual multi-speaker speech corpus for scaling indian TTS](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 68161–68182.
- Aditya Narayan Sankaran, Reza Farahbakhsh, and Noel Crespi. 2026. [Few-shot contrastive adaptation for audio abuse detection in low-resource indic languages](#). *Preprint*, arXiv:2604.09094.
- Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, and 46 others. 2023. [Seamless: Multilingual expressive and streaming speech translation](#). *arXiv preprint arXiv:2312.05187*.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. [SALMONN: Towards generic hearing abilities for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Vincenzo Timmel, Claudio Paonessa, Manfred Vogel, Daniel Perruchoud, and Reza Kakooee. 2025. [Fine-tuning whisper on low-resource languages for real-world applications](#). In *Proceedings of the 10th Edition of the Swiss Text Analytics Conference*, pages 57–65.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. [Representation learning with contrastive predictive coding](#). *arXiv preprint arXiv:1807.03748*.

Laurens van der Maaten and Geoffrey Hinton. 2008. [Visualizing data using t-SNE](#). *Journal of Machine Learning Research*, 9:2579–2605.

A Training and Decoding Details

NE-ASR. Whisper-medium has 307M parameters with a 24-layer encoder and decoder. All parameters are updated during fine-tuning. Training uses fp16 precision, AdamW with weight decay 0.01, linear warmup over 500 steps, and gradient clipping at 1.0. Effective batch size is 32 (batch 16, accumulation 2). Training completes in approximately 4 hours on a single A100 80GB. All zero-shot baselines use greedy decoding (beam size 1), temperature 0.0, and task=transcribe. The same proxy token is used at both training and inference.

NE-SpeechEmbed. Speech projection: Linear(1024, 768). Text projection: Linear(768, 768). Log-temperature initialized at 0.0 (temperature=1.0). Only projection layers and temperature are trained (1,377,793 parameters total). Training completes in approximately 4 hours on a single A100 80GB.

MMS-1B fine-tuning. The backbone is frozen; only adapter layers are trained (50–180K parameters per language). Training uses learning rate 1e-3, batch size 8, gradient accumulation 4, fp16, and 5 epochs per language on a single A100 80GB. For languages without native adapters, the closest available adapter is used as initialization. For Garo, the Mizo (miz) adapter is used due to the Bengali-script mismatch in grt.

Text normalization. All conditions use an identical normalization pipeline before metric computation: (1) lowercasing, (2) removal of noise markers (<noise>, <unk>, <static>), correction annotations ({...}), and bracket markers ([...]), (3) removal of punctuation and the Garo raka interpunct (·, U+00B7), and (4) whitespace normalization. No numeral normalization or diacritic stripping is applied. The pipeline is identical across all languages and all models; Assamese predictions are evaluated in native script without transliteration.

Language	Train	Val	Test	Hrs
Garo	33,480	1,000	1,000	33.21
Assamese	21,589	1,000	1,000	24.71
Chakma	32,807	1,000	1,000	41.50
Kokborok	2,511	274	279	3.47
Mizo	9,850	1,000	1,000	15.40
Nagamese	12,862	1,000	1,000	16.10
Wancho	8,860	1,000	1,000	8.42
Total	121,959	6,274	6,279	142.81

Table 9: Vaani subset split counts.

Language	Prop. Pairs	Vaani Pairs
Khasi	10,000	0
Garo	1,548	8,452
Mizo	5,994	4,006
Nagamese	6,376	3,624
Kokborok	2,105	2,511
Assamese	0	10,000
Wancho	0	8,860
Chakma	0	10,000
Total	26,023	47,453

Table 10: Speech-text pairs for NE-SpeechEmbed training.

B Dataset Split Details

C Garo MMS-1B Script Mismatch: Bengali Outputs

D Language Notes

Nagamese is an Indo-Aryan-based contact creole spoken as a lingua franca in Nagaland, written in Roman script. Its Assamese-derived lexicon with Naga substratum influence produces high speaker-to-speaker variability.

Wancho is a Tibeto-Burman language of Arunachal Pradesh. A dedicated Unicode script exists (2019); our Vaani subset uses Roman transliteration.

Chakma is a Tibeto-Burman language of Mizoram, Tripura, and Arunachal Pradesh. The Chakma abugida script exists; our Vaani subset uses Roman transliteration.

Kokborok is the primary Tibeto-Burman language of Tripura, co-official with Bengali. Both our proprietary and Vaani subsets use Roman script, consistent with widespread everyday writing practice.

Reference (Latin Garo)	MMS-1B (grt) Output	Commentary
ia noksao kosako see tape donaniko nikna gita manenga	ইয়া নকসাও খোসাখো সে তাস্রে দোনানিখো নিকনা গিত্তা মা মিংআ	Phonetically faithful; wrong script
ian gari tarichakram biap ba nok ba dokan onga	য়ান গাডি খারি চাক্রাম বি আ বা নক বা দকখান হ আ	Partial word splits; correct sounds
banga rasinkoba iano nikna gita mana	বাং আ রাসিনখোবা ইয়ানো নিকনা গিত্তা মা মা	Near-perfect phoneme mapping
ia noksao anga mechik saksani chrokaniko nika	ইয়া নকসাও আংআ মেচিক সাকসানি চ্রকানিখো নিক্কা	Script mismatch, structure preserved
iano bolrangkoba nikatna gita manenga	ইয়ানো বলরাংখোবা নিককাংনা গিত্তা মা মেংআ	Consistent Bengali rendering

Figure 4: Full Bengali-script outputs from the MMS-1B grt adapter on five Garo test utterances. The adapter transcribes Latin-script Garo phonetically into Bengali characters, yielding phonetically plausible but orthographically incorrect output. See Table 7 in the main paper for Latin transliterations and commentary.

E Informed Consent Form (English Translation)

The following is the English translation of the informed consent instrument presented to all proprietary corpus participants in their native language prior to recording, administered digitally via Google Forms. Participants could withdraw at any time before data release.

I understand the purpose of this project is to collect speech data to improve AI tools for Northeast Indian languages, benefiting communities in the Northeast and beyond.

I understand that my speech data (all personal details, name and contact information removed) can be released as an Open Source Dataset. I consent to my anonymized recordings being used by researchers and developers to build and improve AI tools. My data will be handled securely, and I retain the right to access or request deletion as per India's Digital Personal Data Protection Act, 2023.