VANUATUASR: OPEN MULTILINGUAL SPEECH RECOGNITION FOR LOW-RESOURCE VANUATU LANGUAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual automatic speech recognition (ASR) remains inaccessible to many languages that lack transcribed speech, digital text, and representation in large-scale pretraining corpora. We present **VanuatuASR**, to our knowledge the first unified multilingual ASR system for Vanuatu, a highly linguistically diverse region where speech and text resources are severely limited. Building on speech curated across seven Vanuatu languages totaling 65.92 hours, we investigate three practical questions: whether joint multilingual fine-tuning can match language-specific recognizers; whether a shared language model trained on pooled transcripts improves decoding when per-language text is too scarce to support individual models; and whether adding untranscribed speech from additional languages benefits self-supervised representation learning. Our system combines SSL on pooled unlabeled speech, language-balanced joint CTC fine-tuning on four transcribed languages, and shared character n-gram LM decoding without language identifiers. Our results show that joint training matches four language-specific models with a single recognizer. The shared LM is the dominant source of improvement, reducing macro WER from 37.37% to 32.12% and remaining effective with as little as 25% of transcribed training data. Expanding the SSL pool with untranscribed speech from three additional languages, however, yields no measurable gain. VanuatuASR outperforms other state-of-the-art baselines under a common evaluation protocol, reducing macro WER by 10.89% relative to the strongest baseline. These results indicate that in severely low-resource settings, transcribing existing recordings and pooling scarce textual resources across languages delivers more value than collecting additional unlabeled speech.

1 INTRODUCTION

Multilingual automatic speech recognition (ASR) trains a single model to transcribe many languages by sharing acoustic and linguistic representations across them. In recent years this approach has advanced rapidly: systems such as XLS-R, Whisper, MMS and OWLS have expanded coverage from tens of languages to hundreds or more, achieving strong recognition on standard benchmarks for many high- and mid-resource languages (Babu et al., 2022; Radford et al., 2023; Pratap et al., 2024; Chen et al., 2025). Yet many low-resource languages remain poorly served, because transcribed speech is scarce, public evaluation data barely exists, and the linguistic structures of these languages are poorly understood. Without targeted development, continued scaling risks widening the gap between languages that already have data and models and those that do not.

Vanuatu makes this gap especially visible. It is one of the most linguistically diverse countries in the world (François et al., 2015). Yet this diversity is scarcely represented in existing speech corpora, pretrained ASR systems, or multilingual benchmarks, leaving Vanuatu languages largely absent from current speech technology. Pacific languages more broadly have received growing attention: recent work has adapted pretrained models to Cook Islands Māori (Coto-Solano et al., 2022) and to individual Vanuatu languages including Bislama, Nafsan and Lelepa (Xiao et al., 2026; Mahmudi et al., 2026), demonstrating that adaptation is feasible even with a few hours of transcribed speech. However, these efforts train a separate recognizer for each language, leaving potential transfer across

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

languages unexplored and not using the limited textual resources available across languages to support decoding.

This omission matters most in the extreme low-resource regime. Modern end-to-end recognizers can learn substantial linguistic regularities from large amounts of paired speech and text. When only a few hours of transcribed speech are available, however, this implicit knowledge is limited. The challenge is particularly acute for Vanuatu languages, which are not represented in the multilingual pretraining data: the model has little opportunity to learn their language-specific orthographic and linguistic patterns either during pretraining or during adaptation. An external LM therefore offers a complementary source of supervision, modeling plausible character sequences from text that need not be paired with speech. In principle, this can improve decoding without requiring additional speech transcription.

We address these questions through **VanuatuASR**, to our knowledge the first unified multilingual ASR system for Vanuatu. We curate transcribed and untranscribed speech collections spanning seven languages of Vanuatu, and systematically study how speech and text resources can be combined under severe data constraints. Rather than proposing a new ASR architecture, we examine three practical design choices:

RQ1. Can a single jointly fine-tuned recognizer across languages match language-specific recognizers across Vanuatu languages?

RQ2. Can a shared character-level LM trained on pooled Vanuatu transcripts improve ASR decoding, particularly as transcribed speech becomes more limited?

RQ3. Does incorporating untranscribed speech from three additional Vanuatu languages during SSL pretraining improve recognition on the target languages?

VanuatuASR outperforms all the baselines, and our results reveal three findings. First, joint multilingual fine-tuning matches four language-specific systems with a single recognizer. Second, a shared character LM trained on pooled transcripts reduces macro WER from 37.37% to 32.12% and remains beneficial with only 25% of the transcribed training data. Thus, text that is insufficient for large scale fine-tuning can still provide useful complementary supervision for LM training. Third, adding untranscribed speech from three related languages during SSL pretraining produces no measurable improvement at this scale, suggesting that labeled supervision rather than acoustic coverage is the binding constraint. Our contributions are:

- **VanuatuASR.** To our knowledge, the first unified multilingual ASR system for Vanuatu, outperforming Whisper, MMS, XLS-R, and OmniASR under a common evaluation protocol and reducing macro WER by 10.89% relative to the strongest baseline.
- **Practical findings for low-resource multilingual ASR.** A systematic study showing that a shared character language model is the most effective component under extreme data scarcity, that joint training incurs no interference cost, and that unlabeled speech does not compensate for the absence of transcribed data, findings applicable to other highly multilingual, under-resourced regions.
- **A curated community-oriented resource.** We document transcribed and untranscribed speech across seven Vanuatu languages. Subject to consent and licenses, we will release model weights, code, documentation, and evaluation resources.

2 Related Work

Multilingual ASR. Multilingual ASR has expanded rapidly in recent years, with systems covering increasingly many of the world languages. Whisper supports recognition in 99 languages (Radford et al., 2023). OWLS investigates model and data scaling with multilingual encoder–decoder models trained on up to 360,000 hours of public speech data spanning 150 languages (Chen et al., 2025). MMS and Omnilingual ASR extend recognition to 1,107 and more than 1,600 languages, respectively (Pratap et al., 2024; Omnilingual ASR Team, 2025). Yet even at this scale, coverage remains highly uneven: the languages included tend to have existing digital text resources and recorded audio, systematically excluding the world's most endangered varieties. Vanuatu illustrates this gap.

Table 1: Corpus statistics by language (ISO 639-3).

Transcribed speech		Untranscribed recordings	
Language	Hours	Language	Hours
Bislama (bis)	13.36	Namakura (nmk)	12.29
Nafsan (erk)	14.71	Lewo (lww)	13.63
Lelepa (lpa)	3.55	Sie (erg)	7.35
Nguna (llp)	1.02		
Total	32.65	Total	33.27

Despite being one of the world's most linguistically diverse regions, its languages are not represented in the supervised training data of these major ASR systems. This reflects a structural data barrier rather than simply a lack of attention: many Vanuatu languages have limited digital presence and little transcribed speech, making them difficult to include in conventional multilingual ASR training. The consequence is particularly acute for language documentation, where manual transcription remains a major bottleneck to recording and preserving endangered speech (Coto-Solano et al., 2022).

ASR for Pacific languages. Pacific ASR has advanced through adapting multilingual pretrained models to languages with limited transcribed speech. Studies on Māori, Hawaiian and Samoan show that low-resource ASR models can be developed from only a few hours of labeled data (D Chaparala et al., 2024; Leoni et al., 2024). However, transcription remains a major bottleneck: the four-hour Cook Islands Māori corpus required three years of community-led work (Coto-Solano et al., 2022). Recent work has also extended ASR to Australian Indigenous languages (Dang et al., 2025; Mylvaganam et al., 2026b;a; Ambikairajah et al., 2025) and Vanuatu languages. A recent study adapts Whisper-small separately to Bislama, Nafsan and Lelepa (Xiao et al., 2026), while Easper develops language-specific ASR models for Vanuatu language documentation (Mahmudi et al., 2026). While these studies establish the feasibility of Vanuatu ASR, they train separate supervised models for each language, limiting transfer from scarce transcribed data across languages. Joint multilingual ASR can instead exploit shared phonological and lexical structure where present, reducing per-language data requirements and providing a more scalable tool for transcription across Vanuatu's linguistically diverse communities.

Language Models for Multilingual Low-Resource ASR. Language models have long complemented acoustic models in ASR by promoting linguistically plausible output sequences. For CTC-based systems which do not explicitly model output-token dependencies, external LMs are commonly used during beam-search decoding or rescoring (Baevski et al., 2020; Pratap et al., 2024). Encoder–decoder models such as Whisper instead embed an implicit LM acquired through large-scale multilingual pretraining (Radford et al., 2023). However, these LMs are poorly adapted to Vanuatu languages, which have little or no representation in multilingual speech and text corpora. Limited paired speech–text data constrains adaptation, while text from any individual language is insufficient to train a reliable language-specific LM. Therefore, training a dedicated multilingual LM on pooled text from the available Vanuatu languages can exploit shared linguistic regularities where present, and guides ASR decoding towards more plausible target-language transcriptions.

3 Vanuatu Speech Corpus

We curate a corpus of transcribed speech and untranscribed recordings from seven Vanuatu languages, sourced from collections archived by the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC).[1] Access to the source materials is governed by PARADISEC's catalogue permissions.

[1] https://catalog.paradisec.org.au/

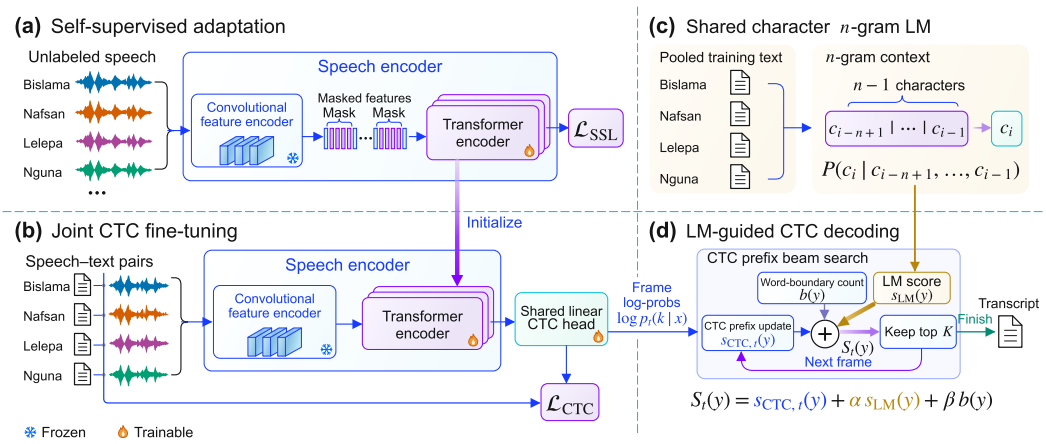

Figure 1: Overview of VanuatuASR. (a) SSL adapts the Transformer encoder on pooled unlabeled speech. (b) Language-balanced joint fine-tuning updates the adapted encoder and shared CTC head. (c) A shared character n-gram LM is trained on pooled transcripts. (d) CTC prefix beam search combines acoustic and LM scores with a word-boundary term to produce a transcript without a language identifier.

Provenance and transcription. The collections arise from long-term language-documentation efforts. The Bislama data comprise interviews and monologues collected by local teams through the Pacific Creoles Project (Thieberger, 2024). The Nafsan collection draws on fieldwork initiated in 1996 (Thieberger, 2012), while the Lelepa archive includes traditional and personal narratives (Lacrampe, 2010). The Nguna recordings were collected in 1978 (Facey, 1978). Community experts prepared transcriptions for these four languages, providing speech–text pairs grounded in community language use and oral narratives.

Language coverage and data preprocessing. Bislama is an English-lexified creole, whereas Nafsan, Lelepa, Nguna, Namakura, Lewo, and Sie belong to the Oceanic branch of Austronesian (Hammarström et al., 2026). The transcribed component contains 24,869 utterances (32.65 h) across Bislama, Nafsan, Lelepa, and Nguna, ranging from 1.02 h for Nguna to 14.71 h for Nafsan. We further curate 33.27 h of untranscribed recordings in Namakura (Sperlich, 2004), Lewo (Early, 2018), and Sie (Crowley, 2006). For the untranscribed recordings, we check audio integrity and divide each recording into non-overlapping 15 s segments. We retain segments with an overall root-mean-square (RMS) level of at least -55 dBFS and at least 20% active frames. This procedure retains 7,990 segments totaling 33.27 h from the original 41.91 h of recordings. Table 1 summarizes both components.

4 Methods

VanuatuASR is a three-stage framework that learns shared representations across all languages in the Vanuatu set. First, we pretrain a Transformer speech encoder using self-supervised learning (SSL) on the pooled unlabeled speech from Vanuatu languages (Figure 1(a)), learning acoustic patterns shared across the set. Next, we fine-tune the encoder on the pooled labeled speech, using language-balanced sampling to mitigate data imbalance (Figure 1(b)). We then train a character-level n-gram LM on the pooled transcripts from all languages to capture shared Vanuatu-specific orthographic and linguistic regularities, and incorporate it during decoding (Figure 1(c)). At inference, acoustic and LM scores are combined to produce transcriptions without requiring the input language identity (Figure 1(d)).

4.1 Encoder Pretraining via SSL

We adapt pretrained speech representations to Vanuatu languages through self-supervised learning (SSL) on a pooled unlabeled Vanuatu speech collection $\mathcal{U}$. This step requires no transcriptions,

allowing the model to learn Vanuatu speech representations from raw audio alone. Pooling unlabeled speech across Vanuatu languages ensures the encoder sees sufficient data to adapt without overfitting to any single language. Concretely, we initialize the acoustic front-end and Transformer encoder from OmniASR CTC 300M v2 (Omnilingual ASR Team, 2025) and continue pretraining with the wav2vec 2.0 contrastive objective (Baevski et al., 2020). For each waveform $\mathbf{x} \in \mathcal{U}$, selected feature spans are masked and passed through the encoder; a quantizer maps the unmasked features to discrete targets (Figure 1(a)). The encoder is trained to match each masked representation u_t to its quantized target q_t rather than sampled distractors:

$$\mathcal{L}_{\text{ctr}} = -\frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \log \frac{\exp(\text{sim}(u_t, q_t)/\tau)}{\sum_{\tilde{q} \in \mathcal{Q}_t} \exp(\text{sim}(u_t, \tilde{q})/\tau)}, \tag{1}$$

where $\mathcal{M}$ denotes masked positions, $\mathcal{Q}_t$ contains the positive target and sampled distractors, and τ is a temperature parameter. The final loss additionally includes the wav2vec 2.0 codebook diversity regularizer. During SSL, the convolutional feature encoder remains frozen, while the Transformer encoder, quantizer, mask embedding and contrastive projections are pre-trained.

4.2 Joint Multilingual CTC Fine-Tuning

To learn a shared multilingual ASR model, we jointly fine-tune a CTC model on the transcribed Vanuatu speech. The model combines the SSL-pretrained encoder and linear CTC head from OmniASR CTC 300M v2 (Figure 1(b)). We retain its tokenizer, whose output vocabulary covers all characters in our normalized transcripts.

Let $p_t(k \mid \boldsymbol{x})$ denote the probability assigned to token k at output frame t for waveform $\boldsymbol{x}$. CTC defines the probability of a tokenized transcription y by summing over all frame-level paths that collapse to y (Graves et al., 2006):

$$P_\theta(y \mid x) = \sum_{\pi : \mathcal{B}(\pi) = y} \prod_{t=1}^{T_x} p_t(\pi_t \mid \boldsymbol{x}), \tag{2}$$

where $\pi = (\pi_1, \ldots, \pi_{T_x})$ is a frame-level path and $\mathcal{B}$ merges repeated tokens and removes blank symbols. We minimize the negative log-likelihood, $\mathcal{L}_{\text{CTC}}(x, y) = -\log P_\theta(y \mid x)$.

The transcribed Vanuatu data are substantially imbalanced across languages (e.g., Nguna with minimal data). To prevent higher-resource languages such as Nafsan from dominating training, we uniformly sample one language for each optimizer update and construct all micro batches for that update from the selected language. Gradients are accumulated across these micro batches, giving each language an equal expected number of updates. Language labels are used only for sampling and are not provided to the recognizer, enabling training of a multilingual ASR.

4.3 Shared character language modeling and decoding

Shared character LM. Vanuatu's languages lack the large text collections needed to train reliable language-specific LMs. We therefore train a single character n gram LM on text pooled across all available Vanuatu languages, sharing the scarce textual signal rather than fragmenting it (Figure 1(c)). Character-level modeling has the additional advantage of scoring words absent from the training transcripts, which is important given the limited vocabulary coverage in small corpora. We treat each normalized utterance as an independent Unicode character sequence with spaces as characters, and estimate each character from up to $n - 1$ preceding characters. Given the limited text available, we apply modified Kneser–Ney smoothing (Chen & Goodman, 1999) to interpolate higher-order estimates with lower-order distributions.

LM-guided CTC decoding. We integrate the LM during CTC prefix beam search (Hannun et al., 2014) (Figure 1(d)). The search extends candidate token prefixes frame by frame, tracking the accumulated CTC log-probability $s_{\text{CTC},t}(y)$ for each prefix y at frame t. To incorporate the LM, each token prefix y is detokenized into a character sequence $c(y)$. The LM then scores this sequence one character at a time, conditioning each character c_i on its preceding context $c_{<i}$:

$$s_{\text{LM}}(y) = \sum_{i=1}^{|c(y)|} \log P_{\text{LM}}(c_i \mid c_{<i}), \tag{3}$$

where the LM conditions on at most $n-1$ preceding characters. As a prefix grows, only the score of the newly added characters is appended, making the update efficient.

At each frame, prefixes are ranked by a combined score that balances acoustic and linguistic evidence:

$$S_t(y) = s_{\mathrm{CTC},t}(y) + \alpha\, s_{\mathrm{LM}}(y) + \beta\, b(y), \tag{4}$$

where α controls the LM contribution and β rewards or penalizes word boundaries, with $b(y) = \max(0, |\,\mathrm{words}(c(y))| - 1)$ counting the number of word boundaries in $c(y)$. The top K prefixes are retained at each frame. After the final frame, an LM end-of-sentence score is added to each surviving prefix, and the highest-scoring normalized transcription is returned.

5 Experimental Setup

Data Preparation. We build our ASR system using the four transcribed languages: Bislama, Nafsan, Lelepa, and Nguna, and three untranscribed speech Vanuatu languages: Namakura, Lewo, and Sie. This gives two SSL pretraining pools: a *default* pool of 26.08 h drawn from the training partitions of the four transcribed languages, and an *expanded* pool that adds 33.27 h of untranscribed speech, totaling 59.35 h. For the transcribed speech, we use 80% for pretraining, 10% for validation and 10% for test. All audio is resampled to 16 kHz. Per-language statistics are reported in Appendix A.

Models and Training. We initialize from OmniASR CTC 300M v2 (Omnilingual ASR Team, 2025) and train in two stages: SSL pretraining on the Vanuatu speech pool, followed by joint CTC fine-tuning on the four transcribed languages. Both stages use AdamW with gradient accumulation over four microbatches per update. SSL is optimized with a learning rate of 10^{-5}. Fine-tuning uses a peak learning rate of 10^{-5} with linear warmup, a constant phase, and exponential decay.

We compare against four pretrained systems. MMS-300M (Pratap et al., 2024) and XLS-R-300M (Babu et al., 2022) are adapted with a randomly initialized CTC head over a shared character vocabulary, first training the head alone, then jointly fine-tuning the contextual encoder and head with the convolutional feature encoder frozen. Both use AdamW at a peak learning rate of 10^{-4} with warmup, a constant phase, and exponential decay. Whisper-small (Radford et al., 2023) is fine-tuned end-to-end with AdamW at a peak learning rate of 10^{-5} with warmup and linear decay. OmniASR fine-tuned without the SSL stage serves as a direct ablation of our pretraining step. OmniASR and Whisper-small are also evaluated without fine-tuning to assess zero-shot transfer. We report test WER and CER averaged over three seeds, with macro-average scores weighting all four languages equally.

Language Modeling and Decoding. We estimate a shared character n-gram LM from the pooled training transcriptions of the four languages using modified Kneser–Ney smoothing, treating each transcription as an independent character sequence with spaces included. Decoding uses CTC prefix beam search with beam width 50 and per-frame acoustic top-k pruning with $k = 8$.

We tune LM order, interpolation weight α, and word insertion penalty β on the development sets. We sweep orders from 4 to 20, and for each order perform a grid search over $\alpha \in \{0.5, 1, 2, 4\}$ and $\beta \in \{-2, -1, -0.5, 0, 0.5, 1, 2\}$, selecting the pair with the lowest mean development WER across the four languages for each seed. A 15-gram LM achieves the lowest average development WER and is used for the final system.

6 Results

6.1 Comparison with open speech recognition systems

Zero-shot transfer fails for Vanuatu languages. Table 2 compares VanuatuASR against open speech models under both zero-shot and fine-tuned conditions on the four target languages. Without fine-tuning, both OmniASR and Whisper produce poor results, with macro WERs of 87.23% and 205.17% respectively. Whisper WER exceeding 200% indicates near-complete failure, confirming that large-scale multilingual pretraining does not transfer to Vanuatu languages and that targeted adaptation is necessary.

Table 2: Word and character error rates (%) of different speech recognition systems on the four target languages

		WER					CER				
System	Params	Bislama	Nafsan	Lelepa	Nguna	Macro	Bislama	Nafsan	Lelepa	Nguna	Macro
Without fine-tuning											
OmniASR	300M	72.62	91.42	99.41	85.45	87.23	28.92	30.58	39.27	23.91	30.67
Whisper	244M	217.58	246.75	244.72	111.62	205.17	201.92	216.71	203.38	58.68	170.17
Four-language joint fine-tuning											
OmniASR	300M	20.41	47.37	62.64	21.67	38.03	**7.99**	14.35	21.00	6.47	12.46
MMS	300M	81.46	90.31	93.28	62.79	81.96	33.50	35.07	38.13	15.79	30.62
XLS-R	300M	21.26	47.20	59.21	**16.51**	36.05	8.62	14.75	20.78	6.05	12.55
Whisper	244M	23.95	56.94	74.34	26.35	45.40	11.22	20.79	29.65	8.25	17.48
This work											
VanuatuASR (no LM)	300M	20.55	46.79	61.22	20.92	37.37	8.00	**14.30**	20.78	6.47	12.39
VanuatuASR	300M	**16.50**	**42.09**	**52.76**	17.13	**32.12**	**7.99**	15.36	**19.77**	**5.77**	**12.22**

Fine-tuning substantially improves all systems. Joint fine-tuning on the four target languages reduces OmniASR's macro WER from 87.23% to 38.03% and Whisper's from 205.17% to 45.40%, demonstrating that even a small amount of locally collected labeled data yields large gains. MMS is the exception: despite being designed for low-resource languages, it remains the weakest fine-tuned system at 81.96% macro WER, suggesting its pretraining representations transfer poorly to this language family.

VanuatuASR achieves the best overall performance. VanuatuASR achieves the lowest macro WER (32.12%) and macro CER (12.22%) among all evaluated systems, outperforming the strongest fine-tuned baseline, XLS-R, by 3.93 WER points. The LM contributes meaningfully: removing it raises macro WER from 32.12% to 37.37%, confirming that the shared character LM provides consistent gains across languages.

6.2 Effectiveness of Language modeling

We further examine whether the shared character LM trained on pooled Vanuatu text provides robust benefits as transcribed training data decreases, a practically important question for endangered language communities where annotation is costly.

The LM provides consistent gains across supervision levels. Table 3 shows that the shared character LM improves WER consistently regardless of how much transcribed data is available, reducing macro WER by 5.15–5.58 points across all conditions. Notably, the absolute WER reduction from the LM is stable even as supervision drops from 100% to 25%, suggesting that the LM's contribution is largely independent of ASR model quality and that pooled textual knowledge from Vanuatu's languages continues to guide decoding toward plausible transcriptions even with minimal data.

The LM partially compensates for reduced supervision. At 25% training data, VanuatuASR with LM achieves 40.55% macro WER, compared to 37.37% for the no-LM system trained on the full data. The LM therefore recovers a substantial portion of the performance lost from reduced supervision, highlighting its practical value in settings where annotation resources are limited and new transcribed data is costly to obtain.

Overall, LM improvements are more pronounced in WER than CER (5.15–5.58 vs 0.17–0.47 points), consistent with a character-level LM correcting word-boundary and lexical errors more than individual character substitutions.

6.3 Joint versus language-specific fine-tuning

A natural question for multilingual ASR is whether training a single joint model introduces interference that harms individual languages, compared to dedicated language-specific models. Ta-

Table 3: Word and character error rates (%) under reduced supervision, macro-averaged across the four target languages. Δ denotes the absolute error reduction over default decoding, in percentage points.

		WER			CER		
Transcribed training data	LM order	Default	With LM	Δ	Default	With LM	Δ
100%	15	37.37	**32.12**	5.25	12.39	**12.22**	0.17
75%	12	38.90	**33.53**	5.37	12.63	**12.28**	0.35
50%	12	41.00	**35.42**	5.58	13.12	**12.65**	0.47
25%	8	45.70	**40.55**	5.15	14.40	**14.21**	0.19

Table 4: Word and character error rates (%) for joint and language-specific fine-tuning with default decoding on the four target languages.

		WER					CER				
Fine-tuning	Models	Bislama	Nafsan	Lelepa	Nguna	Macro	Bislama	Nafsan	Lelepa	Nguna	Macro
Language-specific	4	19.61	46.92	61.09	19.83	36.86	7.58	14.24	20.57	6.31	12.17
Joint	1	20.55	46.79	61.22	20.92	37.37	8.00	14.30	20.78	6.47	12.39

ble 4 addresses this directly by comparing one jointly fine-tuned model against four independently trained models on the same backbone. The jointly trained model achieves the similar macro WER of 37.37% and macro CER of 12.39%, compared to 36.86% and 12.17% for language-specific fine-tuning. These results confirm that no language shows consistent degradation under joint training. This suggests that Vanuatu's languages share sufficient acoustic and linguistic structure that a single model can serve all four without meaningful interference. Practically, joint training reduces the cost of maintaining and deploying four separate models to one, with no performance penalty, making it the clear choice for a low-resource community setting.

6.4 Scaling SSL adaptation

We investigate whether extending SSL pretraining to untranscribed speech from three additional Vanuatu languages (i.e., Namakura, Lewo, and Sie), bringing the total pool from 26.08 h to 59.35 h, improves encoder representations and downstream ASR performance. Despite their diversity, Vanuatu's languages share regional acoustic properties and phonological patterns that could enrich representation learning through joint SSL pretraining (François et al., 2015; Conneau et al., 2021), motivating the question of whether the four-language pool already captures this shared structure or whether additional untranscribed languages can enrich it further. Table 5 shows that the expanded pool yields virtually no change: macro WER shifts from 37.37% to 37.38%, and per-language differences are within expected variance.

This null result carries a practical implication: SSL on the four transcribed languages already captures the shared acoustic structure sufficiently, and at this scale the bottleneck is labeled supervision rather than acoustic coverage. For communities deciding how to allocate limited annotation resources, this offers a concrete recommendation: transcribing existing recordings yields greater benefit than collecting additional untranscribed speech.

6.5 Ablations

Figure 2 isolates the contribution of each component of VanuatuASR by removing SSL adaptation and LM-guided decoding in turn. SSL adaptation alone reduces macro WER from 38.03% to 37.37%, a modest but consistent improvement across three of the four languages. The LM contributes substantially more: adding LM-guided decoding reduces macro WER from 37.37% to 32.12%, a 5.25 point improvement. The gains are consistent across all four languages in WER, confirming that the shared character LM provides reliable benefit regardless of individual language characteristics. CER improvements are more modest (0.17 points), consistent with a character-level LM correcting word-boundary errors more than individual character substitutions. In summary, the

Table 5: Word and character error rates (%) after SSL adaptation on four or seven languages.

		WER					CER				
SSL pool	Hours	Bislama	Nafsan	Lelepa	Nguna	Macro	Bislama	Nafsan	Lelepa	Nguna	Macro
Four languages	26.08	20.55	46.79	61.22	20.92	37.37	8.00	14.30	20.78	6.47	12.39
Seven languages	59.35	20.09	46.50	61.52	21.41	37.38	7.89	14.22	20.88	6.48	12.37

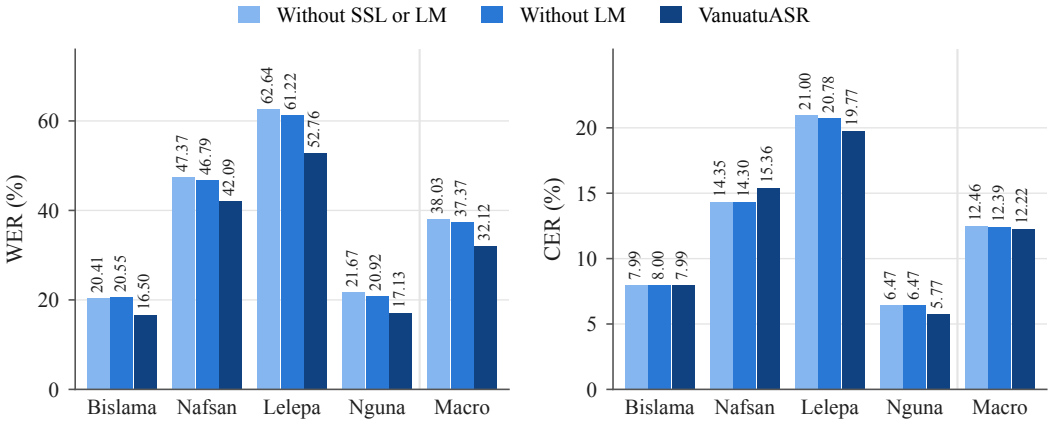

Figure 2: Word and character error rates (%) for component ablations on the four target languages.

LM is the dominant source of improvement, underscoring the value of training a dedicated multilingual language model on pooled Vanuatu text when labeled speech is scarce.

7 Conclusions

We introduced VanuatuASR, the first unified multilingual ASR system for Vanuatu languages, trained on a newly collected corpus of 65.92 hours across seven languages. By combining self-supervised learning on unlabeled Vanuatu speech, joint CTC fine-tuning, and a shared multilingual character language model, our approach pools scarce resources across languages rather than fragmenting them into independent systems. VanuatuASR outperforms globally pretrained systems and language-specific baselines, reducing macro-average WER by 10.89% relative to the strongest baseline.

In summary, Vanuatu's combination of high linguistic diversity, extreme data scarcity and uneven resources across communities is not unique. The same conditions apply across much of other highly multilingual regions with low-resource data. This work, together with the released corpus, model and code, could potentially support continued speech-technology development for the communities absent in the current speech AI landscape.

We also acknowledge that the current system supports four languages for transcription and evaluation. The remaining three languages in the corpus present a natural next step, as expanding transcription coverage will enable broader evaluation and shed light on how well a jointly trained model generalizes to related unseen languages within the same family. Performance across the four supported languages points to a clear direction for improvement: WER ranges from 16.50% for Bislama to 52.76% for Lelepa, a gap that likely reflects differences in training data size and motivates future work on data collection and adaptation strategies more robust to within-set imbalance. Finally, our study is conducted on a single language family in a single region; extending this evaluation framework to other highly multilingual, under-resourced regions would broaden the evidence base for the practical findings reported here and help establish how widely the conclusions generalize.

AI USE STATEMENT

Generative AI tools were used only to polish the language of this manuscript, such as grammar and wording.

REFERENCES

Eliathamby Ambikairajah, Jingyao Wu, Ting Dang, and Vidhyasaharan Sethu. A study of speech embedding similarities between australian aboriginal and high-resource languages. In *Proc. Interspeech 2025*, pp. 1498–1502, 2025.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech*, pp. 2278–2282, 2022. doi: 10.21437/Interspeech.2022-143. URL https://www.isca-archive.org/interspeech_2022/babu22_interspeech.html.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020. URL https://arxiv.org/abs/2006.11477.

Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394, 1999. doi: 10.1006/csla.1999.0128. URL https://doi.org/10.1006/csla.1999.0128.

William Chen, Jinchuan Tian, Yifan Peng, Brian Yan, Chao-Han Huck Yang, and Shinji Watanabe. OWLS: Scaling laws for multilingual speech recognition and translation models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 9121–9145. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25bj.html.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021*, pp. 2426–2430, 2021. doi: 10.21437/Interspeech.2021-329. URL https://www.isca-archive.org/interspeech_2021/conneau21_interspeech.html.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Koka'ua, Syed Tanveer, and Isaac Feldman. Development of automatic speech recognition for the documentation of Cook Islands Māori. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 3872–3882. European Language Resources Association, 2022. URL https://aclanthology.org/2022.lrec-1.412/.

Terence Crowley. Sie (Vanuatu). PARADISEC digital collection, 2006. URL https://doi.org/10.4225/72/56e976d34786d.

Kaavya D Chaparala, Guido Zarrella, Bruce Torres Fischer, Larry Kimura, and Oiwi Parker Jones. Mai Ho'omāuna i ka 'Ai: Language models improve automatic speech recognition in Hawaiian. In *Proceedings of The 4th NeurIPS Efficient Natural Language and Speech Processing Workshop*, volume 262 of *Proceedings of Machine Learning Research*, pp. 576–583. PMLR, 2024. URL https://proceedings.mlr.press/v262/d-chaparala24a.html.

Ting Dang, Trini Manoj Jeyaseelan, Eliathamby Ambikairajah, and Vidhyasaharan Sethu. Characterization of speech similarity between australian aboriginal and high-resource languages: A case study on dharawal. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 658–663. IEEE, 2025.

Robert Early. Epi recordings. PARADISEC digital collection, 2018. URL https://doi.org/10.26278/5b88148ba12bd.

Ellen Facey. Nguna material. Collection EF1, Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), 1978. URL https://doi.org/10.4225/72/56E97A453362F. Ellen Facey (collector).

Alexandre François, Michael Franjieh, Sébastien Lacrampe, and Stefan Schnell. The exceptional linguistic density of Vanuatu. In Alexandre François, Sébastien Lacrampe, Michael Franjieh, and Stefan Schnell (eds.), *The Languages of Vanuatu: Unity and Diversity*, pp. 1–21. Asia-Pacific Linguistics, Canberra, 2015.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 369–376, 2006. doi: 10.1145/1143844.1143891. URL https://www.cs.toronto.edu/~graves/icml_2006.pdf.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. *Glottolog 5.3*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2026. doi: 10.5281/zenodo.18840935. URL https://glottolog.org. Accessed on 15 September 2026.

Awni Y. Hannun, Andrew L. Maas, Daniel Jurafsky, and Andrew Y. Ng. First-pass large vocabulary continuous speech recognition using bi-directional recurrent DNNs. *arXiv preprint arXiv:1408.2873*, 2014. URL https://arxiv.org/abs/1408.2873.

Sebastien Lacrampe. Lelepa texts and recordings (2006–2009). PARADISEC digital collection, 2010. URL https://doi.org/10.4225/72/56e976959cd39.

Gianna Leoni, Lee Steven, Miles Thompson, Tūreiti Keith, Keoni Mahelona, Peter-Lucas Jones, and Suzanne Duncan. Solving failure modes in the creation of trustworthy language technologies. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pp. 325–330, Torino, Italia, 2024. ELRA and ICCL. URL https://aclanthology.org/2024.sigul-1.39/.

Aso Mahmudi, Ting Dang, Ekaterina Vylomova, and Nick Thieberger. Easper: An accessible ASR pipeline for language documentation. *arXiv preprint arXiv:2608.11629*, 2026. URL https://arxiv.org/abs/2608.11629.

Pravina Mylvaganam, Eliathamby Ambikairajah, Ting Dang, Vidhyasaharan Sethu, and Tuende Szalay. Which languages transfer best to warlpiri? a similarity-based study for low-resource asr. *Interspeech*, 2026a.

Pravina Mylvaganam, Ting Dang, Eliathamby Ambikairajah, Vidhyasaharan Sethu, and Jingyao Wu. Hybrid continual learning for low-resource australian aboriginal language identification. *Interspeech*, 2026b.

Omnilingual ASR Team. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*, 2025. URL https://arxiv.org/abs/2511.09690.

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. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024. URL http://jmlr.org/papers/v25/23-1318.html.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Wolfgang Sperlich. Namakir (Namakura), Vanuatu & Niue. PARADISEC digital collection, 2004. URL https://doi.org/10.4225/72/56e9772b627d7.

Nick Thieberger. South Efate (Vanuatu). PARADISEC digital collection, 2012. URL https://doi.org/10.4225/72/56e97595b6d0a.

Nick Thieberger. Bislama corpus. PARADISEC digital collection, 2024. URL https://doi.org/10.26278/6tr4-q919.

Yang Xiao, Aso Mahmudi, Nick Thieberger, Eliathamby Ambikairajah, Eun-Jung Holden, and Ting Dang. Continual adaptation for Pacific Indigenous speech recognition. *arXiv preprint arXiv:2603.06310*, 2026. URL https://arxiv.org/abs/2603.06310.

Table 6: Supervised speech partitions for the four recognition languages. Duration refers to the training partition; counts are numbers of utterances.

Language	Training (h)	Training	Development	Test
Bislama	10.72	9,209	1,150	1,148
Nafsan	11.77	7,379	921	925
Lelepa	2.82	2,416	302	304
Nguna	0.82	892	111	112
Total	26.12	19,896	2,484	2,489

Table 7: Untranscribed audio before and after preparation for SSL.

Language	Original (h)	Retained (h)	SSL segments
Namakura	13.32	12.29	2,951
Lewo	20.60	13.63	3,273
Sie	7.99	7.35	1,766
Total	41.91	33.27	7,990

A Data Statistics

Supervised speech. Table 6 reports the utterance counts and training durations for the four recognition languages. The training partitions contain 19,896 utterances totaling 26.12 h.

Self-supervised speech pools. After SSL-specific filtering, the four-language SSL pool contains 19,702 segments totaling 26.08 h, drawn from the 26.12 h supervised training partitions. Table 7 summarizes the original untranscribed audio and the segments we prepare for SSL from Namakura, Lewo, and Sie. Adding the retained audio expands the pool to 27,692 segments and 59.35 h. Segments in the expanded pool range from 0.605 to 30 s.