PUMA: Projected Universal Multilingual ASR for Low-Resource Settings. Application to Diverse African Languages

Ilyes Oukid[1,3], Bilal Faye[1], Hanane Azzag[1], Mustapha Lebbah[2], Said Yacine Boulahia[3]
[1]LIPN CNRS UMR 7030, USPN, Villetaneuse, France
[2]DAVID Lab, UVSQ, Paris-Saclay University, Versailles, France
[3]Ecole Militaire Polytechnique, Algiers, Algeria
{oukid,faye,azzag}@lipn.univ-paris13.fr
mustapha.lebbah@uvsq.fr, saidyacine.boulahia@emp.mdn.dz

Abstract

Multilingual ASR systems often fail to generalize to low-resource and linguistically diverse languages while remaining costly to scale. We introduce PUMA, a unified multilingual ASR model that improves low-resource performance with reduced model complexity. PUMA employs a Universal Language Projection (ULP) module that integrates a learnable language token with acoustic representations, enabling language-aware processing through shared parameters. Experiments on diverse African languages show consistent word error rate reductions over strong multilingual baselines, highlighting improved robustness and generalization. Our code is available at the following GitHub URL: https://github.com/ilyes-okd/PUMA

1 Introduction

Recent advances in Automatic Speech Recognition (ASR) have been driven by large-scale self-supervised pretraining and transformer-based architectures that effectively model long-range temporal dependencies (Baevski et al., 2021; Liu et al., 2023; Rekesh et al., 2023; Sudo et al., 2024; Xue et al., 2024; Team, 2025). Despite these successes, ASR performance remains poor for low-resource and linguistically diverse languages. This gap is especially pronounced for many African languages, where limited annotated data, rich morphology, and high phonetic variability challenge models trained predominantly on high-resource languages. Multilingual ASR aims to mitigate data scarcity by sharing representations across languages, but existing approaches typically rely on language-specific adapters, conditioning modules, or fine-tuning strategies (Chen et al., 2024; Hu et al., 2024; Inoue et al., 2024; Bai et al., 2024; Samin et al., 2025; Xue et al., 2025). While effective for a fixed set of languages, these designs introduce parameter growth and hinder scalability, making it difficult to extend models to new or underrepresented languages. As a result, current multilingual systems struggle to balance language awareness with efficient parameter sharing, leading to limited generalization in low-resource settings.

We address this limitation by proposing PUMA, a unified multilingual ASR model that achieves language-aware processing *without* introducing language-specific components. PUMA is built around a shared *Universal Language Projection* (ULP) module that operates on acoustic representations fused with learnable language tokens. This design enables the model to encode language-specific cues through a single, shared projection mechanism, preserving scalability while improving cross-lingual transfer.

We evaluate PUMA on a benchmark of low-resource African languages under monolingual, progressive multilingual, and simultaneous multilingual training settings. Across all configurations, PUMA consistently outperforms strong multilingual baselines, demonstrating improved robustness and generalization without increasing model complexity.

The main contributions of this work are as follows:

- We introduce PUMA, a lightweight and scalable multilingual ASR architecture that achieves language-aware transcription without relying on language-specific adapters or task-specific parameters.
- We propose a unified language conditioning mechanism based on a Universal Language Projection (ULP) module with learnable language tokens, enabling full parameter sharing across languages while remaining adaptable to low-resource settings.
- We conduct an extensive evaluation on ten low-resource African languages, covering monolingual, progressive multilingual, and

Findings of the Association for Computational Linguistics: ACL 2026, pages 371–382

simultaneous multilingual training regimes.

- We demonstrate that PUMA achieves competitive or superior performance compared to substantially larger models on both supervised and zero-shot ASR benchmarks, while using significantly fewer trainable parameters.

2 Related Work

Multilingual ASR with Shared Representations. Early end-to-end multilingual ASR systems explored training a single model across multiple languages to learn shared acoustic representations and enable cross-lingual transfer (Toshniwal et al., 2018; Kannan et al., 2019; Pratap et al., 2020; Li et al., 2021; Tjandra et al., 2023). By sharing parameters, these models improve recognition for underrepresented languages through transfer from high-resource data. However, fully shared architectures often struggle to model language-specific phonological and morphological characteristics, particularly for linguistically distant or highly imbalanced languages, leading to degraded performance in low-resource settings such as many African languages.

Self-Supervised Speech Representation Learning. Self-supervised learning has significantly advanced multilingual ASR by enabling the learning of robust acoustic representations from large amounts of unlabeled speech. Models such as wav2vec 2.0 (Baevski et al., 2020), XLSR (Conneau et al., 2021), XLS-R (Babu et al., 2022), and the Universal Speech Model (USM) (Zhang et al., 2023) have demonstrated strong transfer to low-resource languages after supervised fine-tuning. Nevertheless, these approaches typically rely on downstream adaptation and offer limited mechanisms for explicitly modeling language identity, which can hinder robustness when applied to highly diverse or morphologically rich language families.

Adapter-Based Multilingual ASR. To address the limitations of fully shared models, adapter-based multilingual ASR architectures introduce language-specific modules on top of a shared backbone (Yu et al., 2023). Massively Multilingual Speech (MMS) (Pratap et al., 2024), for example, builds on wav2vec 2.0 representations and employs a dedicated adapter per language to improve specialization. While effective, such designs increase parameter count linearly with the number of languages, reducing scalability and complicating deployment in low-resource scenarios where new languages must be added efficiently.

Large-Scale and Zero-Shot Multilingual ASR. Recent work has focused on large-scale and zero-shot ASR to extend coverage to languages unseen during supervised training. Whisper (Radford et al., 2023) achieves broad multilingual coverage through large-scale multitask training, while dedicated zero-shot approaches further target unseen-language recognition (Zhao et al., 2025; Omnilingual et al., 2025). Although these methods demonstrate impressive generalization, they typically require very large model capacity and extensive training resources, making them less suitable for frugal or targeted low-resource ASR applications.

Positioning of PUMA. In contrast to prior work, PUMA targets scalable and low-resource multilingual ASR without introducing language-specific components or increasing parameter count as the number of languages grows. By integrating language information through a shared Universal Language Projection (ULP) mechanism guided by learnable language tokens, PUMA enables language-aware modeling while preserving a compact architecture. This design is particularly well suited to low-resource and linguistically diverse settings, such as African languages, where both data availability and deployment efficiency are critical.

3 Method

We introduce PUMA, a frugal multilingual architecture for automatic speech recognition designed for under-annotated and low-resource languages, with a particular focus on African languages. PUMA relies on full parameter sharing across languages and explicit language conditioning through learnable language tokens, enabling efficient transfer across languages while limiting the number of trainable parameters. This design allows the model to be trained effectively in low-resource settings. Figure 1 illustrates the overall processing pipeline. PUMA is composed of five core components:

1. **Feature Extractor (FE)** transforms raw audio waveforms into low-level acoustic features.

2. **Feature Projection (FP)** projects these features into a fixed-dimensional space and applies normalization to ensure compatibility with subsequent modules.

3. **Universal Language Projection (ULP)** conditions the acoustic representations on the target language using learnable language tokens.

4. **Encoder (En)** captures temporal dynamics and long-range contextual dependencies within the sequence.

5. **Language Modeling Head (LM Head)** maps the encoded representations to the target vocabulary, producing output probability distributions.

3.1 PUMA Architecture

Let a mini-batch of input audio signals be denoted by $\mathbf{x} \in \mathbb{R}^{N \times D_{in}}$, where N is the mini-batch size and D_{in} denotes the dimensionality of the audio representation after preprocessing. The model outputs a sequence of probability distributions $\mathbf{z} \in \mathbb{R}^{N \times L \times M}$, where L is the output sequence length and M is the size of the target vocabulary.

Feature Extractor and Feature Projection. The FE and FP components are derived from a pretrained wav2vec 2.0 encoder (Baevski et al., 2020) and are kept **frozen** during training to reduce computational cost and facilitate transfer learning. The FE extracts intermediate acoustic representations capturing phonetic and prosodic characteristics, which are then projected by FP into a latent space of dimension d:

$$\mathbf{h} = \text{FP}(\text{FE}(\mathbf{x})) \in \mathbb{R}^{N \times L \times D_{fp}}. \quad (1)$$

This step reduces acoustic variability and standardizes representations across languages.

Universal Language Projection (ULP). The ULP module is the core component of PUMA. Its goal is to enable a single shared parameter set, θ, to represent multiple languages without relying on language-specific adapters, as in MMS (Pratap et al., 2024).

In MMS, each language l is associated with a dedicated adapter ϕ_l, and inference is formulated as:

$$\hat{\mathbf{y}}_l = \arg\max_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}_l; \theta, \phi_l), \quad (2)$$

where $\mathbf{x}_l$ denotes the acoustic input sequence for language l, and $\mathbf{y}$ the predicted output token sequence. This design leads to parameter growth proportional to the number of languages, with each adapter introducing over two million additional parameters.

PUMA replaces these adapters with a lightweight learnable language token $\mathbf{t}_l \in \mathbb{R}^{1 \times D_{fp}}$. This token injects language information directly into the acoustic representations while preserving a fully shared parameterization.

For a given language l, the token $\mathbf{t}_l$ is concatenated with the projected features $\mathbf{h}$:

$$\mathbf{h}' = \text{concat}(\mathbf{t}_l, \mathbf{h}) \in \mathbb{R}^{N \times (L+1) \times D_{fp}}. \quad (3)$$

The resulting sequence is processed by $n = 4$ shared Transformer blocks (Vaswani et al., 2017), using RMSNorm (Zhang and Sennrich, 2019):

$$\mathbf{u} = \text{ULP}(\mathbf{h}'; \theta) \in \mathbb{R}^{N \times (L+1) \times D_{fp}}. \quad (4)$$

After processing, the language tokens are removed from the output:

$$\mathbf{u} = \mathbf{u}[:, 1 : L, :]. \quad (5)$$

This is necessary because language tokens do not correspond to actual audio spans; retaining them would disrupt alignment between model outputs and ground-truth transcriptions during training, especially in alignment-sensitive decoding scenarios such as CTC or attention-based loss functions.

Encoder. The ULP output $\mathbf{u}$ is combined with the original acoustic features $\mathbf{h}$ via a residual connection, and the resulting representation is processed by the encoder to model temporal and long-range contextual dependencies:

$$\mathbf{e} = \text{Encoder}(\mathbf{h} + \mathbf{u}) \in \mathbb{R}^{N \times L \times D_{enc}}. \quad (6)$$

The encoder corresponds to the transformer encoder block of a pretrained wav2vec 2.0 model. Its parameters are used to initialize the encoder and are kept **frozen** throughout training to reduce computational cost and enable efficient adaptation to low-resource languages.

Language Modeling Head. A linear projection with weights $W \in \mathbb{R}^{D_{enc} \times M}$ and bias $b \in \mathbb{R}^M$ maps encoder outputs to the vocabulary space. A softmax produces output probabilities:

$$\mathbf{z} = \text{Softmax}(W\mathbf{e} + b) \in \mathbb{R}^{N \times L \times M}. \quad (7)$$

This output $\mathbf{z}$ is used for decoding the final transcription under a sequence modeling objective.

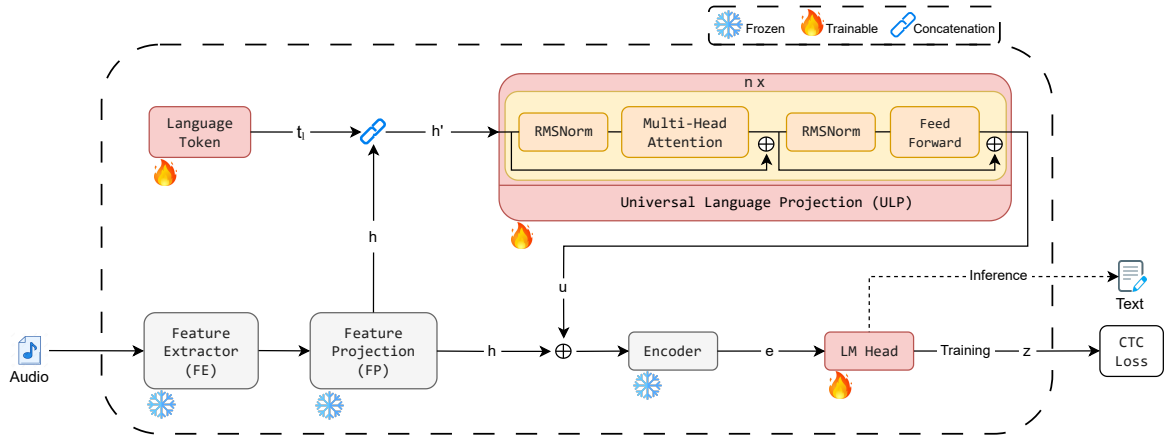

Figure 1: Overview of the PUMA architecture with its five components: Feature Extractor (FE), Feature Projection (FP), Universal Language Projection (ULP), Encoder, and LM Head.

3.2 Training Procedure

PUMA can be trained under three complementary configurations to evaluate robustness and scalability across different multilingual scenarios: monolingual, progressive multilingual, and simultaneous multilingual training.

Monolingual Training. In this configuration, PUMA is trained separately on each language l. Let $\mathcal{D}_l$ denote the training dataset for language l. Training optimizes the objective:

$$\theta_l^* = \arg\min_{\theta} \sum_{(\mathbf{x}_l, \mathbf{y}_l) \in \mathcal{D}_l} \mathcal{L}\big(f_\theta(\mathbf{x}_l, \mathbf{t}_l), \mathbf{y}_l\big), \quad (8)$$

where $\mathbf{t}_l$ is the learnable language token for language l. This setup produces a separate model for each language and serves as a baseline to measure the effect of multilingual training strategies.

Progressive Multilingual Training. In this configuration, languages are introduced sequentially. The model is first trained on an initial language l_0. When a new language l_1 is added, the shared parameters θ are initialized with the weights learned from l_0 and then fine-tuned jointly with the new language token $\mathbf{t}_{l_1}$. To prevent catastrophic forgetting, a small subset of data from previously learned languages is replayed during training. Formally, after k languages have been integrated, the objective becomes:

$$\theta^* = \arg\min_{\theta} \sum_{l \in \{l_0, \ldots, l_k\}} \sum_{(\mathbf{x}_l, \mathbf{y}_l) \in \tilde{\mathcal{D}}_l} \mathcal{L}\big(f_\theta(\mathbf{x}_l, \mathbf{t}_l), \mathbf{y}_l\big), \quad (9)$$

where $\tilde{\mathcal{D}}_l$ denotes either the full dataset for the current language or a small replay set for previous languages. This approach limits interference between languages, stabilizes training, and enables an open-ended system that can incorporate new languages incrementally.

Simultaneous Multilingual Training. Let the multilingual dataset be defined as the union of all languages:

$$\mathcal{D} = \bigcup_{l=1}^{N_l} \mathcal{D}_l, \quad (10)$$

where N_l is the total number of languages. At each training iteration, mini-batches are sampled from $\mathcal{D}$, and the model parameters are optimized by:

$$\theta^* = \arg\min_{\theta} \sum_{(\mathbf{x}_l, \mathbf{y}_l) \in \mathcal{B}} \mathcal{L}\big(f_\theta(\mathbf{x}_l, \mathbf{t}_l), \mathbf{y}_l\big), \quad (11)$$

where $\mathcal{B}$ denotes the sampled mini-batch. This configuration promotes cross-lingual generalization but may lead to gradient interference between languages.

Loss Function. Across all training configurations, we employ the Connectionist Temporal Classification (CTC) loss (Graves et al., 2006) to align predicted sequences with target transcriptions. Given model outputs $\mathbf{z} \in \mathbb{R}^{N \times L \times M}$, the CTC loss is defined as:

$$\mathcal{L}_{\text{CTC}} = \text{CTCLoss}\big(\log(\mathbf{z}), \mathbf{y}, l_{\text{in}}, l_{\text{lab}}\big), \quad (12)$$

where $\log(\mathbf{z})$ are the log-probabilities over the vocabulary, $\mathbf{y}$ the target label sequences, l_{in} the input lengths after encoding, l_{lab} the label lengths, and a special blank token is used to model unaligned time steps. This formulation allows the model to learn an implicit alignment between acoustic representations and target transcriptions without requiring frame-level annotations.

4 Experiments

We evaluate PUMA against strong state-of-the-art multilingual ASR baselines. Performance is assessed using standard evaluation metrics: Word Error Rate (WER) to measure transcription accuracy, as well as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) scores to measure sentence-level alignment and fluency.

4.1 Datasets

We evaluate PUMA on ten low-resource African languages, including Kabyle (TutlaytAI, 2024), Wolof (Gilbert, 2024), Hausa (Global, 2025), Yoruba (Agili, 2025), Igbo (Nwankwo, 2025), Shona (Speech, 2025), Xhosa (Mattingly, 2024), Fon (Suru, 2024), Twi (Sicoli, 2024), and Arabic (Mohamed, 2024). These languages are characterized by limited data availability and substantial linguistic diversity, making them particularly challenging for multilingual ASR systems. On average, each language provides approximately 15 hours of transcribed speech. All audio files are resampled to 16 kHz, and consistent train/validation/test splits are defined across languages. We construct a shared vocabulary by merging character inventories from all languages, resulting in a unified vocabulary of 204 characters. A common tokenizer is then applied to all datasets, ensuring identical text processing across languages.

4.2 Experimental Setup

All experiments are conducted under identical conditions to ensure fair comparison. Models are trained using the AdamW optimizer with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-3}, with an effective batch size of 32 achieved through gradient accumulation. The Universal Language Projection (ULP) encoder consists of 4 Transformer blocks with a hidden dimension of 768, 8 self-attention heads, and a projection dimension of 1280. All experiments are performed on a single NVIDIA A100 GPU (80 GB).

4.3 Training Results

We evaluate PUMA primarily against MMS (Pratap et al., 2024), which is considered a strong state-of-the-art baseline for multilingual ASR. The main objective of this comparison is to assess whether full parameter sharing guided by learnable language tokens can achieve performance comparable to, or better than, MMS, which relies on language-specific adapters.

MMS relies on a backbone with 1B parameters, and its total parameter count increases linearly with the number of supported languages due to the addition of a dedicated adapter for each language. In contrast, PUMA maintains a **fixed** parameter budget of 1B parameters regardless of the number of languages. During training, the feature extractor, feature projection, and encoder components are kept frozen, while only the Universal Language Projection (ULP) module, the learnable language tokens, and the language modeling head are updated, corresponding to approximately **30M** trainable parameters. This design enables efficient multilingual adaptation while avoiding the parameter growth inherent to adapter-based approaches.

In addition to MMS, we also report results for Whisper (Radford et al., 2023) as a large-scale multilingual ASR system. Whisper includes a substantially larger number of trainable parameters compared to PUMA across its small (244 M), medium (769 M), and large (1.5 B) variants, making direct comparisons less balanced in terms of model capacity and computational cost. For this reason, and due to computational constraints, Whisper is fine-tuned only under the simultaneous multilingual training configuration. In the monolingual and progressive training settings, comparisons are therefore restricted to MMS.

4.3.1 Monolingual Training

Table 1 presents monolingual ASR results for MMS and PUMA across all evaluated languages. PUMA consistently outperforms MMS on most languages, with notable WER reductions on Kabyle, Wolof, Hausa, Igbo, Fon, and Twi, while also achieving comparable or higher BLEU and ROUGE scores, indicating stronger generalization in low-resource scenarios. For Arabic, ROUGE scores are not reported, as the presence of diacritics and the lack of explicit segmentation complicate reliable character-level matching during evaluation. Overall, these results validate the architectural choices of PUMA in the monolingual setting. We next evaluate its performance under multilingual training using both progressive and simultaneous strategies.

4.3.2 Progressive Multilingual Training

Table 2 reports results obtained under progressive multilingual training, where languages are introduced incrementally while retaining previously

Lang.	Model	WER ↓	BLEU ↑	R-1 ↑	R-2 ↑	R-L ↑
Kabyle	MMS	0.551	0.277	0.695	0.507	0.694
	PUMA	**0.307**	**0.428**	**0.793**	**0.640**	**0.793**
Wolof	MMS	0.278	0.586	**0.848**	**0.740**	**0.848**
	PUMA	**0.215**	**0.606**	0.825	0.704	0.825
Hausa	MMS	0.143	0.809	0.958	0.933	0.958
	PUMA	**0.029**	**0.942**	**0.975**	**0.956**	**0.975**
Yoruba	MMS	0.297	**0.567**	**0.806**	**0.690**	**0.805**
	PUMA	**0.274**	0.565	0.791	0.673	0.790
Igbo	MMS	0.542	0.277	0.669	0.487	0.666
	PUMA	**0.420**	**0.347**	**0.706**	**0.544**	**0.704**
Shona	MMS	**0.310**	**0.520**	**0.760**	**0.608**	**0.759**
	PUMA	0.313	0.512	0.735	0.586	0.733
Xhosa	MMS	0.336	**0.468**	**0.725**	**0.548**	**0.724**
	PUMA	**0.319**	0.438	0.704	0.520	0.704
Fon	MMS	0.197	0.760	**0.938**	**0.906**	**0.937**
	PUMA	**0.126**	**0.777**	0.913	0.855	0.911
Twi	MMS	0.443	0.340	0.697	0.497	0.693
	PUMA	**0.338**	**0.461**	**0.730**	**0.580**	**0.727**
Arabic	MMS	0.483	**0.291**	–	–	–
	PUMA	**0.454**	0.275	–	–	–

Table 1: Monolingual ASR results comparing MMS and PUMA across all evaluated languages. ROUGE-1, ROUGE-2, and ROUGE-L are abbreviated as R-1, R-2, and R-L, respectively. ROUGE scores are not reported for Arabic due to tokenization issues.

Lang.	Model	WER ↓	BLEU ↑	R-1 ↑	R-2 ↑	R-L ↑
Kabyle	MMS	0.551	0.277	0.695	0.507	0.694
	PUMA	**0.322**	**0.406**	**0.777**	**0.616**	**0.776**
Wolof	MMS	0.278	**0.586**	**0.848**	**0.740**	**0.848**
	PUMA	**0.237**	0.569	0.805	0.675	0.805
Hausa	MMS	0.143	0.809	**0.958**	**0.933**	**0.958**
	PUMA	**0.055**	**0.891**	0.952	0.915	0.952
Yoruba	MMS	0.297	0.567	0.806	0.690	0.805
	PUMA	**0.251**	**0.592**	**0.809**	**0.696**	**0.807**
Igbo	MMS	0.542	0.277	0.669	0.487	0.666
	PUMA	**0.458**	**0.304**	**0.682**	**0.510**	**0.679**
Shona	MMS	0.310	**0.520**	**0.760**	**0.608**	**0.759**
	PUMA	**0.297**	0.516	0.748	0.590	0.746
Xhosa	MMS	0.336	**0.468**	**0.725**	**0.548**	**0.724**
	PUMA	**0.304**	0.458	0.717	0.536	0.716
Fon	MMS	0.197	**0.760**	**0.938**	**0.906**	**0.937**
	PUMA	**0.165**	0.709	0.889	0.815	0.888
Twi	MMS	0.443	0.340	0.697	0.497	0.693
	PUMA	**0.285**	**0.484**	**0.782**	**0.627**	**0.780**
Arabic	MMS	0.483	**0.291**	–	–	–
	PUMA	**0.467**	0.258	–	–	–

Table 2: Progressive multilingual ASR results comparing MMS and PUMA across all evaluated languages. ROUGE-1, ROUGE-2, and ROUGE-L are abbreviated as R-1, R-2, and R-L, respectively. ROUGE scores are not reported for Arabic due to tokenization issues.

seen languages in the training data. This strategy helps mitigate catastrophic forgetting, although it requires maintaining stable optimization across heterogeneous language distributions. Under this setting, PUMA consistently outperforms MMS in terms of WER across all languages, while maintaining competitive BLEU and ROUGE scores. Notably, languages such as Yoruba, Shona, Xhosa, and Twi benefit from progressive training compared to the monolingual setting, suggesting effective cross-lingual knowledge transfer. These results confirm that full parameter sharing combined with language-token conditioning remains effective when languages are added progressively, without relying on language-specific adapters.

4.3.3 Simultaneous Multilingual Training

When all languages are trained jointly (Table 3), PUMA achieves the lowest average WER among all compared models, while also obtaining the highest ROUGE scores and maintaining a competitive BLEU score. This suggests that PUMA generalizes more robustly across languages, likely due to its language-conditioned representations introduced by the ULP module, which mitigates interference between languages in a shared model.

Model	WER ↓	BLEU ↑	R-1 ↑	R-2 ↑	R-L ↑
Whisper-small	0.594	0.312	0.612	0.440	0.609
Whisper-medium	0.402	0.439	0.706	0.551	0.704
Whisper-large	0.339	**0.499**	0.745	0.606	0.743
MMS	0.419	0.411	0.743	0.587	0.741
PUMA	**0.314**	0.474	**0.766**	**0.622**	**0.764**

Table 3: Average multilingual results under simultaneous training. ROUGE-1, ROUGE-2, and ROUGE-L are abbreviated as R-1, R-2, and R-L, respectively. Per-language results are provided in Appendix A.

While Whisper-large shows strong performance, MMS and Whisper-medium obtain lower results, and Whisper-small struggles in low-resource settings. Detailed per-language results are provided in Appendix A. Overall, these findings confirm that PUMA scales effectively to simultaneous multilingual training, preserving strong performance across diverse low-resource languages.

4.4 Zero-Shot Evaluation on FLEURS

To further validate our model, we evaluate PUMA on unseen data from the FLEURS benchmark (Conneau et al., 2023), covering Wolof, Hausa, Yoruba, Igbo, Shona, Xhosa, and Arabic—languages observed during training but evaluated

on a different corpus, enabling a cross-corpus zero-shot evaluation.

PUMA is compared against several state-of-the-art multilingual ASR systems, including Whisper (Radford et al., 2023) (small, medium, and large), MMS (Pratap et al., 2024), Seamless-M4T v2 Large (Team, 2025), and Omnilingual ASR (Omnilingual et al., 2025) in four configurations (CTC-1B, CTC-7B, LLM-1B, and LLM-7B).

As shown in Table 4, PUMA demonstrates competitive performance relative to large multilingual baselines. It consistently outperforms Whisper variants, MMS, and Seamless-M4T v2 Large across all reported metrics, despite a substantial disparity in the number of trainable parameters. PUMA achieves performance close to Omnilingual ASR CTC-1B, which slightly outperforms PUMA while relying on 1B trainable parameters. The largest Omnilingual ASR variants, particularly Omni-LLM-7B, achieve the strongest overall results; however, this comes at the cost of significantly increased architectural complexity and computational requirements.

Figure 2 highlights the substantial gap in the number of trainable parameters between PUMA and existing multilingual ASR systems, showing that competitive—and often superior—performance can be achieved with a significantly smaller and fixed parameter budget. PUMA relies on a constant number of trainable parameters (approximately 30M), regardless of the number of supported languages, as adding a new language only requires learning a lightweight language token. While MMS also updates around 30M parameters in our setting with ten languages, this number scales linearly as new languages are added due to its reliance on language-specific adapters. In contrast, Whisper, Seamless-M4T, and Omnilingual ASR involve substantially larger numbers of trainable parameters, ranging from 244M to 7B, which is orders of magnitude higher than PUMA.

Overall, these results highlight the strong generalization ability of PUMA under distribution shifts and unseen utterances, despite its compact and frugal design. Detailed per-language results are provided in Appendix B.

4.5 Ablation Study

Table 5 presents an ablation study assessing the impact of removing language tokens from the ULP module under a simultaneous multilingual training setup. The ablated model shows lower performance than the original architecture. Nevertheless, it retains comparable performance on a subset of languages such as Wolof, Igbo, Shona, and Xhosa. This observation highlights the strength of the shared attention-based ULP architecture, which can capture useful cross-lingual regularities even without explicit language conditioning.

However, performance degrades noticeably for more linguistically distant and challenging low-resource languages. In particular, for Arabic, removing language tokens leads to a relative WER increase of 12.1% and a BLEU decrease of 11.3%, reflecting increased cross-lingual interference. These results emphasize the importance of explicit language conditioning in multilingual ASR. Overall, they confirm that language tokens $\mathbf{t}_l$ act as routing vectors, guiding shared representations toward language-specific subspaces,

Model	WER ↓	BLEU ↑	R-1 ↑	R-2 ↑	R-L ↑
Whisper-small	0.846	0.071	0.132	0.045	0.130
Whisper-medium	0.813	0.097	0.176	0.073	0.172
Whisper-large	0.792	0.115	0.199	0.091	0.194
MMS	0.482	0.315	0.564	0.387	0.561
Seamless-M4T	0.652	0.269	0.294	0.118	0.284
Omni-CTC-1B	0.372	0.425	0.690	0.518	0.685
Omni-CTC-7B	0.328	0.480	0.730	0.570	0.726
Omni-LLM-1B	0.323	0.502	0.731	0.582	0.728
Omni-LLM-7B	**0.274**	**0.553**	**0.766**	**0.628**	**0.762**
PUMA	0.407	0.364	0.667	0.488	0.661

Table 4: Average zero-shot ASR performance on the FLEURS benchmark. ROUGE-1, ROUGE-2, and ROUGE-L are abbreviated as R-1, R-2, and R-L, respectively. Seamless-M4T results are averaged over supported languages only. Per-language results are reported in Appendix B.

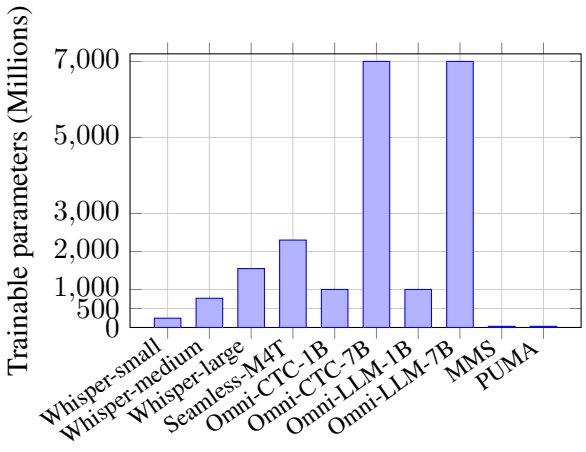

Figure 2: Comparison of the number of trainable parameters between baseline models and PUMA.

thereby mitigating cross-lingual interference while enabling efficient knowledge transfer across typologically diverse languages.

Lang.	Lang. Token	WER ↓	BLEU ↑	R-1 ↑	R-2 ↑	R-L ↑
Kabyle	Present	**0.503**	**0.196**	**0.638**	**0.431**	**0.637**
	Removed	0.548	0.180	0.630	0.424	0.629
Wolof	Present	**0.320**	**0.459**	**0.737**	**0.578**	**0.736**
	Removed	0.335	0.421	0.697	0.507	0.697
Hausa	Present	**0.075**	**0.855**	**0.934**	**0.885**	**0.933**
	Removed	0.095	0.824	0.916	0.858	0.916
Yoruba	Present	**0.245**	**0.588**	**0.828**	**0.714**	**0.826**
	Removed	0.297	0.560	0.783	0.655	0.780
Igbo	Present	**0.503**	**0.246**	**0.655**	**0.474**	**0.652**
	Removed	0.505	0.243	0.643	0.455	0.641
Shona	Present	**0.321**	**0.486**	**0.727**	**0.560**	**0.725**
	Removed	0.330	0.424	0.699	0.508	0.699
Xhosa	Present	**0.329**	**0.422**	**0.695**	**0.506**	**0.695**
	Removed	0.337	0.416	0.692	0.499	0.693
Fon	Present	**0.207**	**0.637**	**0.870**	**0.785**	**0.868**
	Removed	0.296	0.544	0.815	0.694	0.812
Twi	Present	**0.257**	**0.510**	**0.807**	**0.664**	**0.806**
	Removed	0.272	0.508	0.793	0.648	0.792
Arabic	Present	**0.384**	**0.345**	–	–	–
	Removed	0.505	0.232	–	–	–

Table 5: Ablation results of PUMA. *Present* denotes the configuration with language tokens, while *Removed* denotes the configuration without language tokens. ROUGE-1, ROUGE-2, and ROUGE-L are abbreviated as R-1, R-2, and R-L, respectively. ROUGE scores are not reported for Arabic due to tokenization issues.

5 Conclusion

In this work, we introduced PUMA, a frugal and scalable multilingual ASR model built on a unified architecture that integrates a Universal Language Projection (ULP) module combined with learnable language tokens. This design enables language-aware processing through fully shared parameters, allowing efficient multilingual adaptation without language-specific adapters or parameter growth. Extensive experiments demonstrate that PUMA consistently outperforms strong multilingual baselines, despite relying on significantly lower architectural and computational complexity.

The proposed ULP-based architecture facilitates effective cross-lingual transfer and yields robust generalization, particularly for low-resource languages where existing multilingual ASR systems often struggle. Language tokens play a central role in guiding the ULP module, mitigating cross-lingual interference, and preserving performance on linguistically distant and challenging low-resource languages.

As future work, we plan to explore multi-token language representations instead of a single language token, in order to capture richer linguistic characteristics such as phonetic, orthographic, and prosodic traits. This direction could further enhance language-aware processing and improve transcription quality. We also aim to investigate Mixture-of-Experts (MoE) strategies to better integrate diverse languages, potentially improving specialization, scalability, and efficiency in massively multilingual ASR systems.

Limitations

While PUMA demonstrates strong performance across multiple multilingual ASR settings, particularly on low-resource languages, its zero-shot performance remains constrained by its deliberately lightweight and frugal design. In contrast, large-scale multilingual systems such as Omnilingual ASR (Omnilingual et al., 2025) leverage massive model capacity, contextual learning, and LLM-based representations to achieve strong zero-shot performance across diverse languages. A promising direction for future work is to explore frugal LLM-inspired architectures and richer contextual representations within a constrained parameter budget, aiming to improve zero-shot performance while preserving the scalability advantages of lightweight multilingual ASR models, and thereby achieving a better trade-off between frugality and zero-shot capability.

Ethical Considerations

This work does not raise any ethical issues. Our study relies exclusively on publicly available datasets, all used in accordance with their respective licenses and strictly for research purposes. We also build upon open-source pretrained models, which were used as baselines and leveraged in compliance with their intended academic and non-commercial use.

References

Hidi Agili. 2025. Yoruba tts dataset. https://huggingface.co/datasets/Hidi-agili/yoruba_tts_dataset.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, and 1 others. 2022. Xls-r: Self-supervised cross-

lingual speech representation learning at scale. In *Proc. Interspeech 2022*, pages 2278–2282.

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021. Unsupervised speech recognition. *Advances in Neural Information Processing Systems*, 34:27826–27839.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Junwen Bai, Bo Li, Qiujia Li, Tara N Sainath, and Trevor Strohman. 2024. Efficient adapter finetuning for tail languages in streaming multilingual asr. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10841–10845. IEEE.

Yaqi Chen, Hao Zhang, Xukui Yang, Wenlin Zhang, and Dan Qu. 2024. Meta-adapter for self-supervised speech models: A solution to low-resource speech recognition challenges. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11215–11221.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

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. IEEE.

Yiga Gilbert. 2024. Alffa wolof asr dataset. https://huggingface.co/datasets/yigagilbert/alffa-wolof-asr-dataset-19hr.

CLEAR Global. 2025. Hausa synthetic asr dataset (xtts). https://huggingface.co/datasets/CLEAR-Global/Hausa-Synthetic-ASR-Dataset-XTTS.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Qing Hu, Yan Zhang, Xianlei Zhang, Zongyu Han, and Xiuxia Liang. 2024. Language fusion via adapters for low-resource speech recognition. *Speech Communication*, 158:103037.

Nakamasa Inoue, Shinta Otake, Takumi Hirose, Masanari Ohi, and Rei Kawakami. 2024. Elp-adapters: Parameter efficient adapter tuning for various speech processing tasks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Anjuli Kannan, Arindrima Datta, Tara N Sainath, Eugene Weinstein, Bhuvana Ramabhadran, Yonghui Wu, Ankur Bapna, Zhifeng Chen, and Seungji Lee. 2019. Large-scale multilingual speech recognition with a streaming end-to-end model. In *Proc. Interspeech 2019*, pages 2130–2134.

Bo Li, Ruoming Pang, Tara N Sainath, Anmol Gulati, Yu Zhang, James Qin, Parisa Haghani, W Ronny Huang, Min Ma, and Junwen Bai. 2021. Scaling end-to-end models for large-scale multilingual asr. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1011–1018. IEEE.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81. Association for Computational Linguistics.

Alexander H Liu, Wei-Ning Hsu, Michael Auli, and Alexei Baevski. 2023. Towards end-to-end unsupervised speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 221–228. IEEE.

William Mattingly. 2024. Xhosa merged audio dataset. https://huggingface.co/datasets/wjbmattingly/xhosa_merged_audio.

Yahya Mohamed. 2024. Arabic audio rev3 9643 2021 dataset. https://huggingface.co/datasets/Yahya-Mohamed/Arabic_Audio_Rev3_9643_2021_Dataset.

Theo Nwankwo. 2025. Igbo tts normalized dataset. https://huggingface.co/datasets/Twelve2five/igbo_tts_normalized.

ASR Omnilingual, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, and 1 others. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. 2020. Massively multilingual asr: 50 languages, 1 model, 1 billion parameters. *arXiv preprint arXiv:2007.03001*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi,

and 1 others. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, and 1 others. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Ahnaf Mozib Samin, Shekhar Nayak, Andrea De Marco, and Claudia Borg. 2025. Investigating adapters for parameter-efficient low-resource automatic speech recognition. In *Proceedings of the 10th Workshop on Representation Learning for NLP (RepL4NLP-2025)*, pages 100–107.

Fabio Sicoli. 2024. Twi dataset. https://huggingface.co/datasets/fsicoli/twi.

Realtime Speech. 2025. Shona dataset. https://huggingface.co/datasets/realtime-speech/shona1.

Yui Sudo, Muhammad Shakeel, Yosuke Fukumoto, Yifan Peng, and Shinji Watanabe. 2024. Contextualized automatic speech recognition with attention-based bias phrase boosted beam search. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10896–10900. IEEE.

Jonathan Suru. 2024. Fon tts dataset. https://huggingface.co/datasets/jonathansuru/fon_tts.

SEAMLESS Communication Team. 2025. Joint speech and text machine translation for up to 100 languages. *Nature*, 637(8046):587–593.

Andros Tjandra, Nayan Singhal, David Zhang, Ozlem Kalinli, Abdelrahman Mohamed, Duc Le, and Michael L Seltzer. 2023. Massively multilingual asr on 70 languages: Tokenization, architecture, and generalization capabilities. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Shubham Toshniwal, Tara N Sainath, Ron J Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao. 2018. Multilingual speech recognition with a single end-to-end model. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4904–4908. IEEE.

TutlaytAI. 2024. Kabyle asr dataset. https://huggingface.co/datasets/TutlaytAI/kabyle_asr.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Hongfei Xue, Kaixun Huang, Zhikai Zhou, Shen Huang, Shidong Shang, and Lei Xie. 2025. The teaaslp system for multilingual conversational speech recognition and speech diarization in mlc-slm 2025 challenge. In *Proc. MLCSLM 2025*, pages 14–17.

Hongfei Xue, Qijie Shao, Kaixun Huang, Peikun Chen, Jie Liu, and Lei Xie. 2024. Sshr: Leveraging self-supervised hierarchical representations for multilingual automatic speech recognition. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Zhongzhi Yu, Yang Zhang, Kaizhi Qian, Cheng Wan, Yonggan Fu, Yongan Zhang, and Yingyan Celine Lin. 2023. Master-asr: achieving multilingual scalability and low-resource adaptation in asr with modular learning. In *International Conference on Machine Learning*, pages 40475–40487. PMLR.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, and 1 others. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

Jinming Zhao, Vineel Pratap, and Michael Auli. 2025. Scaling a simple approach to zero-shot speech recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

A Per-language Results for Simultaneous Multilingual Training

Table 6 reports detailed per-language ASR results under the simultaneous multilingual training setting. Results are provided for all evaluated languages and compared across Whisper (small, medium, and large), MMS, and PUMA. These results enable a finer-grained analysis of model behavior across individual low-resource languages.

Overall, PUMA achieves strong and stable performance, outperforming MMS as well as Whisper-small and Whisper-medium across all evaluated languages, and surpassing Whisper-large on most languages. For a small subset of languages, such as Shona and Xhosa, Whisper-large slightly outperforms PUMA, while results on Fon and Twi remain highly competitive. Notably, these results are

Lang.	Model	WER ↓	BLEU ↑	R-1 ↑	R-2 ↑	R-L ↑
Kabyle	Whisper-small	0.890	0.081	0.366	0.185	0.363
	Whisper-medium	0.724	0.157	0.453	0.250	0.449
	Whisper-large	0.561	0.273	0.566	0.389	0.565
	MMS	0.619	0.217	0.619	0.404	0.619
	PUMA	**0.503**	**0.196**	**0.638**	**0.431**	**0.637**
Wolof	Whisper-small	0.539	0.255	0.540	0.345	0.537
	Whisper-medium	0.373	0.406	0.677	0.497	0.676
	Whisper-large	0.349	0.437	0.707	0.533	0.704
	MMS	0.405	0.415	0.734	0.571	0.733
	PUMA	**0.320**	**0.459**	**0.737**	**0.578**	**0.736**
Hausa	Whisper-small	0.177	0.704	0.846	0.752	0.846
	Whisper-medium	0.136	0.763	0.886	0.812	0.886
	Whisper-large	0.113	0.809	0.908	0.851	0.908
	MMS	0.199	0.713	0.912	0.847	0.912
	PUMA	**0.075**	**0.855**	**0.934**	**0.885**	**0.933**
Yoruba	Whisper-small	0.364	0.485	0.713	0.562	0.710
	Whisper-medium	0.285	0.575	0.780	0.646	0.778
	Whisper-large	0.272	**0.604**	0.797	0.679	0.794
	MMS	0.339	0.507	0.776	0.637	0.773
	PUMA	**0.245**	0.588	**0.828**	**0.714**	**0.826**
Igbo	Whisper-small	0.743	0.115	0.445	0.240	0.442
	Whisper-medium	0.599	0.235	0.601	0.418	0.597
	Whisper-large	0.613	0.211	0.575	0.384	0.572
	MMS	0.586	0.241	**0.658**	**0.480**	**0.655**
	PUMA	**0.503**	**0.246**	0.655	0.474	0.652
Shona	Whisper-small	0.446	0.382	0.636	0.455	0.635
	Whisper-medium	0.392	0.437	0.704	0.533	0.702
	Whisper-large	**0.305**	**0.508**	**0.742**	**0.580**	**0.741**
	MMS	0.353	0.461	0.721	0.549	0.718
	PUMA	0.321	0.486	0.727	0.560	0.725
Xhosa	Whisper-small	0.432	0.343	0.600	0.399	0.599
	Whisper-medium	0.322	0.457	0.690	0.502	0.690
	Whisper-large	**0.262**	**0.539**	**0.749**	**0.583**	**0.747**
	MMS	0.368	0.422	0.692	0.500	0.692
	PUMA	0.329	0.422	0.695	0.506	0.695
Fon	Whisper-small	0.402	0.453	0.709	0.552	0.705
	Whisper-medium	0.239	0.661	0.823	0.719	0.821
	Whisper-large	**0.200**	**0.709**	0.854	0.771	0.852
	MMS	0.398	0.475	0.801	0.674	0.797
	PUMA	0.207	0.637	**0.870**	**0.785**	**0.868**
Twi	Whisper-small	0.543	0.291	0.651	0.466	0.646
	Whisper-medium	0.327	0.479	0.743	0.584	0.741
	Whisper-large	0.266	**0.565**	**0.807**	**0.684**	0.805
	MMS	0.345	0.476	0.778	0.624	0.774
	PUMA	**0.257**	0.510	**0.807**	0.664	**0.806**
Arabic	Whisper-small	1.403	0.009	–	–	–
	Whisper-medium	0.626	0.218	–	–	–
	Whisper-large	0.446	0.337	–	–	–
	MMS	0.579	0.182	–	–	–
	PUMA	**0.384**	**0.345**	–	–	–

Table 6: Per-language results for simultaneous multilingual training. ROUGE-1, ROUGE-2, and ROUGE-L are abbreviated as R-1, R-2, and R-L, respectively. ROUGE scores are not reported for Arabic due to tokenization issues.

obtained despite the substantial difference in architectural and computational cost, with PUMA relying on a much smaller and more efficient model design.

B Per-language Zero-Shot Results on FLEURS

This section reports detailed per-language zero-shot ASR results on the FLEURS benchmark. We provide fine-grained comparisons across Whisper (small, medium, and large), MMS, Seamless-M4T v2 Large, Omnilingual ASR in four configurations (CTC-1B, CTC-7B, LLM-1B, and LLM-7B), and PUMA for each individual language.

All models are evaluated on languages that are available in FLEURS and were already observed during training, enabling a cross-corpus zero-shot evaluation. Per-language results for Wolof and Hausa are reported in Table 7, results for Yoruba and Igbo in Table 8, and results for Shona, Arabic, and Xhosa in Table 9.

These per-language results enable a more detailed analysis of model generalization across diverse low-resource languages. Despite its compact and frugal architecture, PUMA achieves strong and competitive performance compared to large state-of-the-art multilingual ASR systems.

	Wolof					Hausa				
Model	WER ↓	BLEU ↑	R-1 ↑	R-2 ↑	R-L ↑	WER ↓	BLEU ↑	R-1 ↑	R-2 ↑	R-L ↑
Whisper-small	0.919	0.009	0.153	0.054	0.149	0.944	0.004	0.107	0.033	0.104
Whisper-medium	0.907	0.012	0.175	0.067	0.168	0.928	0.008	0.129	0.046	0.127
Whisper-large	0.905	0.013	0.175	0.070	0.172	0.905	0.014	0.173	0.078	0.169
MMS	0.497	0.246	0.529	0.303	0.525	0.264	0.557	0.767	0.621	0.765
Omni-CTC-1B	0.412	0.354	0.638	0.440	0.633	0.264	0.559	0.779	0.640	0.777
Omni-CTC-7B	0.393	0.359	0.645	0.445	0.640	0.238	0.590	0.805	0.673	0.803
Omni-LLM-1B	0.372	0.405	0.677	0.491	0.673	0.229	0.615	0.816	0.692	0.814
Omni-LLM-7B	**0.347**	**0.418**	**0.689**	**0.504**	**0.685**	**0.199**	**0.645**	**0.826**	**0.708**	**0.824**
PUMA	0.426	0.337	0.617	0.424	0.614	0.324	0.465	0.721	0.552	0.718

Table 7: Zero-shot ASR performance on the FLEURS benchmark for Wolof and Hausa. ROUGE-1, ROUGE-2, and ROUGE-L are abbreviated as R-1, R-2, and R-L, respectively.

	Yoruba					Igbo				
Model	WER ↓	BLEU ↑	R-1 ↑	R-2 ↑	R-L ↑	WER ↓	BLEU ↑	R-1 ↑	R-2 ↑	R-L ↑
Whisper-small	0.971	0.002	0.086	0.017	0.082	0.944	0.003	0.117	0.033	0.113
Whisper-medium	0.966	0.003	0.130	0.035	0.122	0.915	0.010	0.184	0.073	0.176
Whisper-large	0.963	0.003	0.138	0.040	0.131	0.897	0.012	0.226	0.110	0.217
MMS	0.841	0.089	0.227	0.104	0.219	0.652	0.143	0.460	0.238	0.455
Seamless-M4T	0.772	0.087	0.348	0.152	0.334	0.929	0.059	0.249	0.083	0.235
Omni-CTC-1B	0.528	0.254	0.697	0.550	0.685	0.547	0.249	0.638	0.445	0.630
Omni-CTC-7B	0.505	0.283	0.697	0.550	0.682	0.475	0.321	0.704	0.531	0.698
Omni-LLM-1B	0.516	0.308	0.644	0.518	0.635	0.456	0.360	0.711	0.551	0.705
Omni-LLM-7B	**0.460**	**0.352**	**0.697**	**0.579**	**0.689**	**0.364**	**0.453**	**0.750**	**0.604**	**0.746**
PUMA	0.626	0.157	0.616	0.451	0.600	0.467	0.311	0.674	0.488	0.665

Table 8: Zero-shot ASR performance on the FLEURS benchmark for Yoruba and Igbo. ROUGE-1, ROUGE-2, and ROUGE-L are abbreviated as R-1, R-2, and R-L, respectively.

	Shona					Arabic		Xhosa				
Model	WER ↓	BLEU ↑	R-1 ↑	R-2 ↑	R-L ↑	WER ↓	BLEU ↑	WER ↓	BLEU ↑	R-1 ↑	R-2 ↑	R-L ↑
Whisper-small	0.875	0.032	0.212	0.103	0.212	0.334	0.439	0.936	0.007	0.120	0.031	0.119
Whisper-medium	0.834	0.043	0.268	0.148	0.268	0.240	0.583	0.899	0.017	0.173	0.068	0.172
Whisper-large	0.829	0.043	0.280	0.163	0.279	0.163	0.695	0.883	0.023	0.200	0.086	0.199
MMS	0.245	0.568	0.770	0.622	0.769	0.472	0.249	0.401	0.353	0.634	0.432	0.634
Seamless-M4T	0.826	0.078	0.286	0.119	0.284	**0.080**	**0.853**	–	–	–	–	–
Omni-CTC-1B	0.235	0.582	0.785	0.643	0.784	0.172	0.674	0.446	0.304	0.603	0.388	0.602
Omni-CTC-7B	0.183	0.657	0.832	0.709	0.832	0.143	0.724	0.357	0.428	0.700	0.514	0.700
Omni-LLM-1B	0.193	0.646	0.823	0.700	0.823	0.142	0.736	0.353	0.441	0.717	0.539	0.717
Omni-LLM-7B	**0.150**	**0.719**	**0.865**	**0.764**	**0.864**	0.122	0.762	**0.279**	**0.521**	**0.766**	**0.610**	**0.765**
PUMA	0.255	0.548	0.764	0.612	0.763	0.330	0.420	0.420	0.313	0.609	0.400	0.608

Table 9: Zero-shot ASR performance on the FLEURS benchmark for Shona, Arabic, and Xhosa. ROUGE-1, ROUGE-2, and ROUGE-L are abbreviated as R-1, R-2, and R-L, respectively. ROUGE scores are not reported for Arabic due to tokenization issues. Seamless-M4T does not support Xhosa.