Efficient ASR for Low-Resource Languages: Leveraging Cross-Lingual Unlabeled Data

Srihari Bandarupalli[1*], Bhavana Akkiraju[1*], Charan Devarakonda[1],
Vamsiraghusimha Narsinga[1], Anil Kumar Vuppala[1]
[1]Speech Processing Lab, International Institute of Information Technology Hyderabad, India
{srihari.bandarupalli,bhavana.akkiraju,sricharan.d}@research.iiit.ac.in
narasinga.vamshi@research.iiit.ac.in, anil.vuppala@iiit.ac.in
* These authors contributed equally.

Abstract

Automatic speech recognition for low-resource languages remains fundamentally constrained by the scarcity of labeled data and computational resources required by state-of-the-art models. We present a systematic investigation into cross-lingual continuous pretraining for low-resource languages, using Perso-Arabic languages (Persian, Arabic, and Urdu) as our primary case study. Our approach demonstrates that strategic utilization of unlabeled speech data can effectively bridge the resource gap without sacrificing recognition accuracy. We construct a 3,000-hour multilingual corpus through a scalable unlabeled data collection pipeline and employ targeted continual pretraining combined with morphologically-aware tokenization to develop a 300M parameter model that achieves performance comparable to systems 5 times larger. Our model outperforms Whisper Large v3 (1.5B parameters) on Persian and achieves competitive results on Arabic and Urdu despite using significantly fewer parameters and substantially less labeled data. These findings challenge the prevailing assumption that ASR quality scales primarily with model size, revealing instead that data relevance and strategic pretraining are more critical factors for low-resource scenarios. This work provides a practical pathway toward inclusive speech technology, enabling effective ASR for underrepresented languages without dependence on massive computational infrastructure or proprietary datasets. To encourage further research, we release our entire codebase and model checkpoints[1].

1 Introduction

Accurate ASR for morphologically complex, low-resource languages remains a pressing challenge. Recent self-supervised methods like Wav2Vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) have advanced ASR using large-scale unlabeled data. Building on this, multilingual models like XLS-R (Conneau et al., 2021) have improved cross-lingual ASR with diverse pretraining (Wang et al., 2021; Pratap et al., 2020; Ardila et al., 2020; Valk and Alumäe, 2020; Cui et al., 2015). However, these gains mainly benefit high-resource languages, while low-resource ones still suffer from limited labeled data.

For languages like Persian, Arabic, and Urdu, the intersection of limited labeled data, complex orthographies, and rich morphology poses considerable difficulties for ASR. Recent systematic evaluations of state-of-the-art models for Persian[2], Arabic (Wang et al., 2024), and Urdu (Arif et al., 2025) indicate that models like Whisper Large v3 (Radford et al., 2023) and Seamless Large v2 (Communication et al., 2023) exhibit superior performance. However, they are computationally intensive (1.5B+ parameters), making efficient deployment challenging. Some prior works (Getman et al., 2024) have explored continuous pretraining, but their analysis focuses on high-resource scenarios and single-language improvements.

In this work, we address these gaps by systematically investigating ASR for low-resource languages, using Perso-Arabic languages as our primary case study. Our approach combines scalable unlabeled data collection with targeted continuous pretraining and morphologically-aware tokenization to demonstrate that compact models (300M parameters) can achieve performance comparable to systems five times larger. Through systematic evaluation across Persian, Arabic, and Urdu, we provide empirical evidence that strategic utilization of cross-lingual unlabeled data can effectively overcome resource constraints without sacrificing recognition accuracy.

[1] https://github.com/sriharib128/EfficientASR.git

[2] https://huggingface.co/spaces/navidved/open_persian_asr_leaderboard

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1611–1617
December 20-24, 2025 ©2025 Association for Computational Linguistics

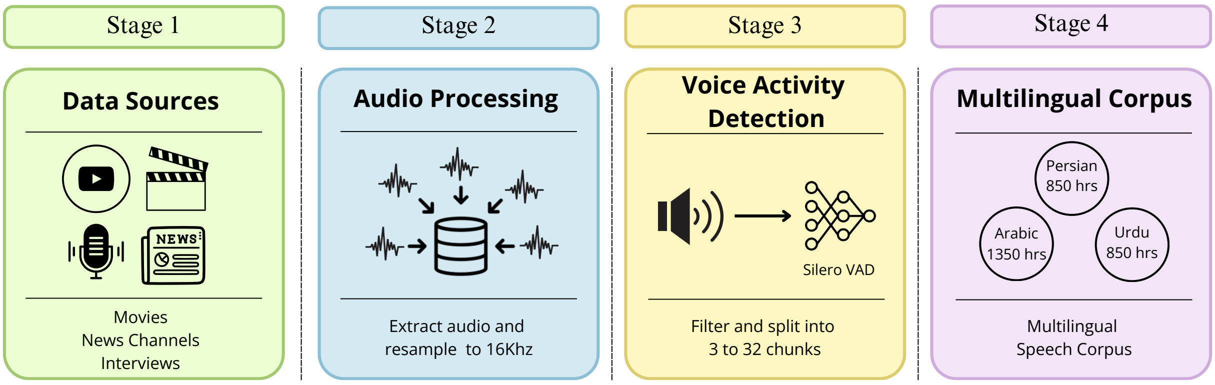

Figure 1: Systematic pipeline for constructing a robust, multilingual unlabeled speech corpus.

Language	Pretraining	Train Set	Test Set
Urdu	816 hrs	60 hrs	10 hrs
Persian	878 hrs	69 hrs	11 hrs
Arabic	1,310 hrs	74 hrs	11 hrs

Table 1: Data statistics showing pretraining corpus (after VAD-based segmentation) and labeled train/test splits.

2 Data

We systematically constructed both unlabeled and labeled speech corpora using a modular data infrastructure to enable controlled investigation of cross-lingual ASR for morphologically-complex, low-resource languages.

2.1 Unlabeled Corpus for Robust Pretraining

We developed a scalable pipeline (Figure 1) to curate high-quality, domain-diverse unlabeled speech for cross-lingual self-supervised training. We acquired multimedia content (films, news broadcasts, interviews) from publicly accessible resources for Persian, Arabic, and Urdu, with language verification through platform metadata or manual inspection. Audio tracks were extracted, resampled to 16 kHz, and processed using Silero VAD[3] to remove non-speech segments, retaining only chunks with speech probability exceeding 70% (Ramirez et al., 2024). Audio was segmented into 3-32 second chunks suitable for training, resulting in the pretraining corpus detailed in Table 1.

2.2 Labeled Data

We curated evaluation-ready labeled corpora by combining data from multiple sources to remove bias and ensure diversity across speakers, dialects, and contexts. Sources include Common Voice (Ardila et al., 2020), and other datasets:

- **Urdu**: IndicVoices (Javed et al., 2024), Urdu Speech-To-Text Dataset[4]
- **Persian**: Persian Speech Corpus[5], ParsiGoo[6], Persian Speech[7], ManaTTS-Persian-Speech-Dataset (Qharabagh et al., 2024), Persian text-to-speech audio (Moradi, 2024)
- **Arabic**: OpenSLR (Kolobov et al., 2021), MGB-2 (Ali et al., 2016)

All audio was resampled to 16 kHz with normalized transcription following (Bandarupalli et al., 2025). Data was stratified into training and test sets as shown in Table 1. Detailed breakdowns of the training and validation splits for each dataset are provided in Appendix A.

3 Experiments

We design systematic experiments to assess efficient automatic speech recognition (ASR) for low-resource, morphologically complex Persian-Arabic languages. The experimental design is guided by three research questions: *(1) Can cross-lingual unlabeled speech improve ASR performance in low-resource Perso-Arabic languages? (2) Which pretraining initialization strategy yields the best adaptation with continual pretraining? (3) Can compact, parameter-efficient models achieve performance comparable to much larger state-of-the-art systems through strategic pretraining and tokenization?* Figure 2 summarizes our training pipeline.

3.1 Pretraining Strategies

To answer these questions, we compare three distinct pretraining initialization strategies:

[3] github.com/snakers4/silero-vad.git

[4] www.kaggle.com/datasets/themohal/tiny-urdu-speech-text-dataset

[5] fa.persianspeechcorpus.com

[6] huggingface.co/datasets/Kamtera/ParsiGoo

[7] github.com/persiandataset/PersianSpeech

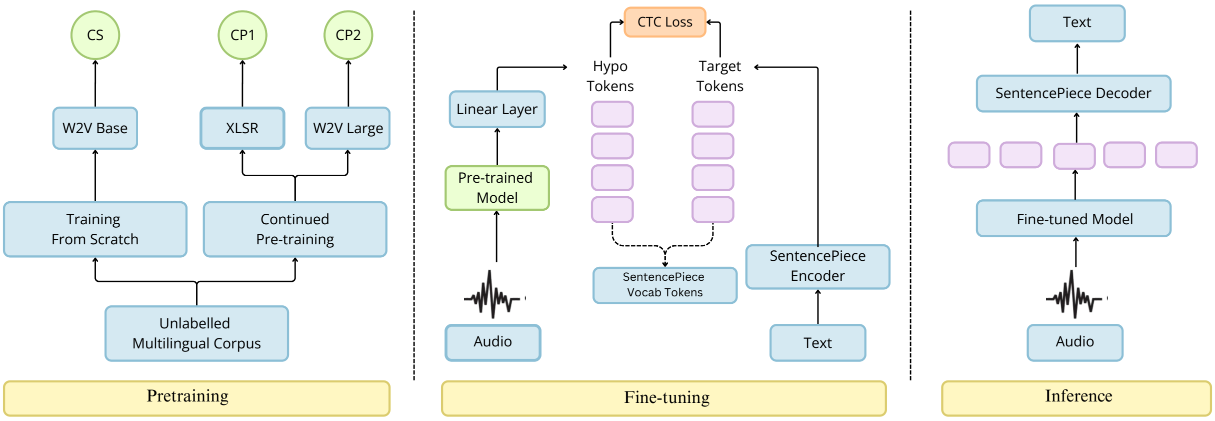

Figure 2: Overview of our experimental framework. The pipeline illustrates three training strategies: CS (Wav2Vec 2.0 Base trained from scratch), CP1 (XLS-R 300M with continuous pretraining), and CP2 (Wav2Vec 2.0 Large with continuous pretraining). All models undergo pretraining on our 3,000-hour multilingual corpus followed by language-specific fine-tuning with SentencePiece tokenization.

From-Scratch Training (CS): Wav2Vec 2.0 Base (95M parameters), trained entirely from randomly initialized weights on our 3,000-hour multilingual corpus. This setup isolates the effect of cross-lingual unlabeled data without prior representation learning.

Cross-lingual Continual Pretraining (CP1): XLS-R (300M parameters), pretrained on 436K hours across 128 languages. We further adapt it to our multilingual corpus to examine whether broad multilingual exposure is beneficial for Perso-Arabic languages.

English-centric Continual Pretraining (CP2): Wav2Vec 2.0 Large (300M parameters), pretrained on 65K hours of primarily English speech. This setting tests whether pretraining on a high-resource language can transfer effectively to the target language family.

All three configurations use identical fine-tuning data and hyperparameters to isolate the effect of initialization strategy.

3.2 Tokenization with SentencePiece

Perso-Arabic languages exhibit orthographic complexity and rich morphology, making character-level tokenization inadequate. To address this, we integrate **SentencePiece** with Byte-Pair Encoding (BPE) into the Wav2Vec 2.0 framework. **Language-specific** subword vocabularies are learned from training transcripts and used to initialize the CTC output layer, enabling direct subword prediction. During inference, subword sequences are decoded into words for final transcription. Further implementation details are provided in Appendix B. Empirical comparisons, in Table 7, confirm that SentencePiece-based tokenization yields reductions in WER compared to character-level decoding.

3.3 Training Procedure

All models are trained in two stages: (1) *unsupervised pretraining* on the unlabeled corpus, followed by (2) *supervised fine-tuning* on the labeled datasets. Hyperparameters and train/validation splits are standardized across all settings.

3.3.1 Pretraining

The CS model is trained from scratch for 200k steps, requiring substantially longer optimization to learn meaningful representations. In contrast, CP1 and CP2 leverage pretrained representations and converge within 40k continual pretraining steps. Across all conditions, early stopping based on validation loss is applied to avoid overfitting during pretraining.

3.3.2 Fine-tuning

Language	CS	CP1	CP2
Urdu	55.8	41.1	43.2
Persian	28.5	21.4	19.9
Arabic	71.4	49.8	58.9

Table 2: Validation loss post fine-tuning.

Each model is fine-tuned separately for Persian, Arabic, and Urdu for up to 50k steps with early stopping. While WER is our primary evaluation metric, we additionally report **validation loss** (Table 2) to capture optimization dynamics. Validation loss offers complementary insight into convergence and stability across languages, highlighting that

Model	Urdu	Persian	Arabic
Our Pretrained Models			
CS (from scratch)	25.0	25.4	38.1
CP1 (XLS-R init.)	22.2	22.3	35.3
CP2 (W2V Large init.)	20.6	17.1	32.9
Baseline Models			
W2V Large (300 M)	39.3	33.2	48.7
Seamless Large v2 (2.3B)	31.9	41.1	34.8
Whisper Large v3 (1.5B)	17.2	21.4	27.2

Table 3: Comprehensive WER comparison across all evaluated models after fine-tuning on labeled data.

CP1 and CP2 converge substantially faster and to lower minima compared to CS. Final evaluation is performed using WER on held-out test sets (Table 3).

3.4 Baseline Fine-tuning

To contextualize our results, we benchmark against three strong baselines: *Whisper Large v3*, *Seamless Large v2*, and *Wav2Vec 2.0 Large*. These baselines are fine-tuned only on our labeled datasets without additional pretraining on our unlabeled corpus. Whisper and Seamless represent state-of-the-art systems (1.5B+ and 2.3B parameters, respectively), while Wav2Vec 2.0 Large serves as a capacity-matched control.

For fairness, all baselines were fine-tuned until convergence using publicly available scripts, with early stopping triggered once validation loss indicated overfitting. This matches the stopping criteria applied to our models.

Full hyperparameter details, along with our training scripts, are released on github for reproducibility. All experiments were conducted on **NVIDIA A100 80GB GPUs**.

4 Results and Analysis

We systematically evaluate our approaches to address the three core research questions, presenting quantitative results in Table 3.

4.1 Impact of Cross-Lingual Pretraining Data

The comparison between Wav2Vec 2.0 (W2V) Large and our three models (CS, CP1, CP2) highlights the central role of in-domain unlabeled data. When fine-tuned directly without additional pretraining, W2V Large (300M parameters, 65K hours of English exposure) yields substantially higher WERs across all languages. In contrast, our CS model—trained from scratch on only 3K hours of Perso-Arabic unlabeled audio with 95M parameters—already outperforms W2V Large. This result underscores that domain-relevant pretraining, even at smaller scale, provides greater benefit than model size or large but typologically distant corpora. Both CP1 and CP2 further reduce WERs, confirming the value of targeted continual pretraining.

4.2 Initialization Strategy Analysis

Our three initialization strategies exhibit distinct performance trends. As expected, CS, trained from scratch, underperforms relative to CP1 and CP2 due to its lack of prior knowledge and smaller capacity. More interestingly, CP2 consistently outperforms CP1 despite its smaller pretraining corpus. We hypothesize that this difference stems from the distribution of pretraining data: CP2 was initialized from ∼65K hours of English-only data, whereas CP1 was initialized from ∼65K hours of English combined with ∼375K hours of predominantly European languages. Since pretraining fundamentally shapes the acoustic feature extractor, English-only pretraining may yield representations that are more broadly transferable whereas, when the pretraining corpus is heavily dominated by distant languages, the learned representations may become tuned to phonetic patterns common in those distant languages, thereby leading to a negative transfer to target languages. In other words, while pretraining is generally beneficial up to a point, excessively broad exposure to unrelated language families may hinder performance compared to more targeted pretraining.

4.3 Resource Efficiency

Our 300M parameter CP2 model demonstrates remarkable efficiency while maintaining competitive performance against much larger state-of-the-art systems. Most notably, our model outperforms Whisper Large v3 on Persian despite using 5× fewer parameters, and achieves performance within 3-6% WER of Whisper Large v3 for Urdu and Arabic while maintaining significant parameter advantage

The performance gap can be attributed to Whisper's extensive prior exposure to labeled data—739 hours for Arabic and 104 hours for Urdu—compared to our model's training exclusively on our smaller curated dataset. For Persian, where Whisper had minimal prior exposure (24 hours), our targeted pretraining strategy demonstrates clear advantages.

This comparison demonstrates two critical points: (1) effective use of unlabeled data can compensate for limited labeled supervision, and (2) parameter efficiency does not preclude competitive or even superior performance. By exploiting targeted continual pretraining, CP2 provides a low-resource pathway to high-quality ASR, contrasting with the underperformance of much larger systems such as Seamless Large v2 (2.3B parameters), which failed to converge under the same conditions. These findings establish that model scale alone is insufficient—adaptation strategy and data relevance are key to success in low-resource ASR.

5 Conclusion

In this work, we present a systematic investigation into resource-efficient ASR for morphologically complex, low-resource languages, with a focus on Perso-Arabic script languages as a representative case study. Our 300M parameter model achieves results competitive with state-of-the-art systems over 5× larger, using substantially less labeled data and computational resources.

Our findings challenge the prevailing assumption that ASR quality scales primarily with model size and data volume. Instead, we show that targeted cross-lingual continuous pretraining, morphologically-aware tokenization, and careful data curation are more critical factors for low-resource languages. Comparing our continual pretraining variants reveals that the relevance of pretraining data, not merely its scale, drives effective transfer learning. This insight suggests that focused, linguistically-informed approaches may be more valuable than broad multilingual exposure for underrepresented languages.

Through our scalable data curation pipeline and strategic utilization of unlabeled multilingual data, we provide a practical pathway toward high-quality ASR that is independent of massive proprietary datasets or computational infrastructure. This work contributes to a more inclusive vision of speech technology where linguistic diversity is not limited by resource constraints, enabling robust ASR systems for the hundreds of millions of speakers of low-resource languages worldwide.

Limitations

Our study presents several limitations. First, while we show that cross-lingual pretraining benefits individual low-resource languages, our fine-tuning is performed separately for each language. We have not yet explored the effects of multilingual joint fine-tuning, which could further improve robustness and transfer. Second, although our unlabeled speech corpus is carefully curated for quality and diversity, it may not encompass the full range of dialects, domains, and spontaneous speech present in real-world scenarios. Third, our evaluation is centered on academic benchmark datasets, leaving open questions regarding the practical robustness and end-user acceptance of our models in deployment settings.

References

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. pages 279–284.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Samee Arif, Aamina Jamal Khan, Mustafa Abbas, Agha Ali Raza, and Awais Athar. 2025. WER we stand: Benchmarking Urdu ASR models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5952–5961, Abu Dhabi, UAE. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Srihari Bandarupalli, Bhavana Akkiraju, Sri Charan Devarakonda, Harinie Sivaramasethu, Vamshiraghusimha Narasinga, and Anil Vuppala. 2025. Towards unified processing of Perso-Arabic scripts for ASR. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 23–28, Abu Dhabi, UAE. Association for Computational Linguistics.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, and 46 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *Preprint*, arXiv:2312.05187.

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

Jia Cui, Brian Kingsbury, Bhuvana Ramabhadran, Abhinav Sethy, Kartik Audhkhasi, Xiaodong Cui, Ellen Kislal, Lidia Mangu, Markus Nussbaum-Thom, Michael Picheny, Zoltan Tüske, Pavel Golik, Ralf Schlüter, Hermann Ney, Mark J. F. Gales, Kate M. Knill, Anton Ragni, Haipeng Wang, and Phil Woodland. 2015. Multilingual representations for low resource speech recognition and keyword search. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 259–266.

Yaroslav Getman, Tamas Grosz, and Mikko Kurimo. 2024. What happens in continued pre-training? analysis of self-supervised speech models with continued pre-training for colloquial finnish asr. In *Interspeech 2024*, pages 5043–5047.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP:1–1.

Tahir Javed, Janki Nawale, Eldho George, Sakshi Joshi, Kaushal Bhogale, Deovrat Mehendale, Ishvinder Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, Sneha Ravishankar, Saranya Sukumaran, Tripura Panchagnula, Sunjay Murali, Kunal Gandhi, Ambujavalli R, Manickam M, C Vaijayanthi, Krishnan Karunganni, and 2 others. 2024. IndicVoices: Towards building an inclusive multilingual speech dataset for Indian languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10740–10782, Bangkok, Thailand. Association for Computational Linguistics.

Rostislav Kolobov, Olga Okhapkina, Olga Omelchishina, Andrey Platunov, Roman Bedyakin, Vyacheslav Moshkin, Dmitry Menshikov, and Nikolay Mikhaylovskiy. 2021. Mediaspeech: Multilanguage asr benchmark and dataset. *Preprint*, arXiv:2103.16193.

Nima Moradi. 2024. Persian text-to-speech audio.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, pages 2757–2761.

Mahta Fetrat Qharabagh, Zahra Dehghanian, and Hamid R. Rabiee. 2024. Manatts persian: a recipe for creating tts datasets for lower resource languages. *Preprint*, arXiv:2409.07259.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. ICML'23. JMLR.org.

Francis McCann Ramirez, Luka Chkhetiani, Andrew Ehrenberg, Robert McHardy, Rami Botros, Yash Khare, Andrea Vanzo, Taufiquzzaman Peyash, Gabriel Oexle, Michael Liang, Ilya Sklyar, Enver Fakhan, Ahmed Etefy, Daniel McCrystal, Sam Flamini, Domenic Donato, and Takuya Yoshioka. 2024. Anatomy of industrial scale multilingual ASR. *CoRR*, abs/2404.09841.

Jörgen Valk and Tanel Alumäe. 2020. Voxlingua107: A dataset for spoken language recognition. *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Yingzhi Wang, Anas Alhmoud, and Muhammad Alqurishi. 2024. Open universal arabic asr leaderboard. *Preprint*, arXiv:2412.13788.

A Labeled Data Distribution Across Datasets

Most existing studies evaluating Automatic Speech Recognition (ASR) models focus on assessing performance using a single dataset. For instance, models trained on the Common Voice dataset are typically evaluated only on the same dataset. However, when these models are tested on other datasets without fine-tuning, their performance often proves suboptimal, highlighting a lack of generalizability.

To address this limitation and improve the robustness of ASR models across diverse datasets, we adopted a multi-dataset approach for training and evaluation. For each target language, we combined a fraction of data from multiple datasets, including Common Voice, to create comprehensive training and testing splits. This approach ensures that models are exposed to a wider variety of speech patterns, accents, and recording conditions, thereby enhancing their generalizability.

Below, we provide the exact durations of the datasets used for each language, along with the specific train and validation splits, as summarized in Tables 4, 5, and 6.

Urdu Dataset	Train Duration (hours)	Validation Duration (hours)
Common Voice	4.23	0.73
SME_news	7.48	1.28
Tiny Urdu Speech Corpus	5.41	0.95
Indic Voices	42.9	7.70

Table 4: Urdu dataset splits for fine-tuning.

Persian Dataset	Train Duration (hours)	Validation Duration (hours)
Common Voice	36.5	6.44
Persian Speech Corpus	2.28	0.21
My Audio Tiny	2.25	0.34
TTS Female	22.6	4.04
Moradi	0.83	0.13
ParsiGOO	3.56	0.64

Table 5: Persian dataset splits for fine-tuning.

B SentencePiece Tokenization

To improve the tokenization process in our speech recognition pipeline, we incorporated SentencePiece tokenization within the Fairseq framework. We first trained a Byte-Pair Encoding (BPE) SentencePiece model with a vocabulary size of 512, using the transcriptions from the training dataset. This vocabulary was subsequently utilized to initialize the Connectionist Temporal Classification (CTC) layer of the wav2vec model.

B.1 Integration of SentencePiece in Fairseq

In the default character-based tokenization used in Fairseq, sentences are split into individual characters. To integrate SentencePiece, we modified this process by first segmenting the sentence into words. Instead of directly splitting words at the character level, we applied SentencePiece encoding to tokenize each word into subword units. This approach retains meaningful subword structures while reducing the token sequence length.

During inference, the decoded tokens were grouped at the word level, and each set of tokens was converted back into corresponding words. Finally, the words were concatenated to reconstruct the complete transcription.

B.2 Impact on Word Error Rate (WER)

To evaluate the impact of SentencePiece tokenization, we fine-tuned our CP2 model using both the default character-based tokenization and the proposed SentencePiece-based approach. The WER comparison across Urdu, Persian, and Arabic is summarized in Table 7.

The results demonstrate a significant reduction in WER across all three languages, highlighting the effectiveness of subword-level tokenization over character-based tokenization in CTC-based speech recognition.

Arabic Dataset	Train Duration (hours)	Validation Duration (hours)
Common Voice	27.5	4.89
SLR-108	8.5	1.53
MGB	37.0	5.11

Table 6: Arabic dataset splits for fine-tuning.

Model	Urdu	Persian	Arabic
CP2 (Character-based)	25.8	26.2	39.0
CP2 (SentencePiece-based)	20.6	17.1	32.9

Table 7: Word Error Rate (WER) comparison between character-based and SentencePiece-based tokenization.