Benchmarking Whisper for Low-Resource Speech Recognition: An N-Shot Evaluation on Pashto, Punjabi, and Urdu

Najm Ul Sehar[1], Ayesha Khalid[1], Farah Adeeba[2], Sarmad Hussain[1]
[1]Center for Language Engineering, KICS, UET, Lahore, Pakistan
[2]Department of Computer Science, UET, Lahore, Pakistan
{najm.sehar, ayesha.khalid, sarmad.hussain}@kics.edu.pk,
farah.adeeba@uet.edu.pk

Abstract

Whisper, a large-scale multilingual model, has demonstrated strong performance in speech recognition benchmarks, but its effectiveness on low-resource languages remains underexplored. This paper evaluates Whisper's performance on Pashto, Punjabi, and Urdu, three underrepresented languages. While Automatic Speech Recognition (ASR) has advanced for widely spoken languages, low-resource languages still face challenges due to limited data. Whisper's zero-shot performance was benchmarked and then its small variant was fine-tuned to improve transcription accuracy. Significant reductions in Word Error Rate (WER) were achieved through few-shot fine-tuning, which helped the model better handle challenges such as complex phonetic structures, compared to zero-shot performance. This study contributes to improving multilingual ASR for low-resource languages and highlights Whisper's adaptability and potential for further enhancement.

1 Introduction

The globalization of technology and communication increasingly necessitates the development of effective natural language processing (NLP) tools for low-resource languages. These languages, spoken by millions, are often underrepresented in computational linguistics. Languages such as Pashto, Punjabi, and Urdu play vital roles in diverse cultural contexts, yet their development for ASR is hampered by a scarcity of labeled data (Krasadakis et al., 2024) and limited computational resources. As a result, existing ASR systems struggle to provide accurate solutions, limiting access to critical technologies in areas like voice-activated devices, education, healthcare, and government services.

Zero-shot learning, which allows models to perform tasks on languages they were not explicitly trained for, has emerged as a promising solution (Yang et al., 2024). OpenAI's Whisper (Radford et al., 2023), a transformer-based ASR model, benefits from large-scale multilingual data, enabling strong performance across multiple languages even without language-specific fine-tuning. However, while zero-shot models generalize effectively, their performance on low-resource languages is hindered (Waghmare et al., 2023) by the lack of sufficient training data and an inability to capture unique phonetic, morphological, and syntactic features, resulting in lower transcription accuracy.

Languages like Pashto, with unique phonological structures, require fine-tuning on language-specific datasets for optimal accuracy (Sher et al., 2024). Fine-tuning pre-trained models like Whisper has been shown to improve ASR performance in low-resource settings, reducing WER even with limited data (Liu and Qu, 2024; Pratama and Amrullah, 2024; Do et al., 2023a). Few-shot fine-tuning, using as little as four hours of data, has demonstrated resource efficiency and adaptability, achieving near-optimal performance (Talafha et al., 2023).

Benchmarking ASR systems on multilingual datasets has become a focal point of recent research (Maheshwari et al., 2024). While Whisper's performance on languages like Urdu has been explored in prior studies (Arif et al., 2024), Pashto and Punjabi have not yet been evaluated in this context.

This study addresses this gap and Whisper's zero-shot ASR performance on Pashto, Punjabi, and Urdu was benchmarked and the impact of few-shot fine-tuning on language-specific datasets was assessed. Results show that few-shot fine-tuning significantly improves Whisper's performance, emphasizing the importance of domain-specific adaptation for better ASR accuracy in low-resource settings.

2 Dataset and Preprocessing

Datasets for Pashto, Punjabi, and Urdu were curated to capture linguistic variations and speaker demographics for few-shot fine-tuning and evaluat-

ing the Whisper model on low-resource languages. Details of these datasets are provided below.

2.1 Pashto Dataset

For experimentation, the ELRA-S0381 Dataset[1] is used which includes 108 hours of transcribed broadcast news in Standard Afghan Pashto from over 1,000 speakers across five sources, such as Ashna TV[2], Azadi Radio[3], Deewa Radio [4], Mashaal Radio[5] and Shamshad TV[6]. This dataset, with 46,000 segments and 1.1 million words, provides a robust foundation for Pashto ASR. For this study, a carefully selected 15-hour dataset from 300 speakers was used for n-shot learning, while 4.8 hours from 137 speakers were reserved for evaluation, ensuring diverse accents and age groups.

2.2 Punjabi Dataset

The lack of any publicly available dataset for the Majhi dialect of Punjabi as spoken in Pakistan, along with its corresponding Shahmukhi annotation, necessitated the creation of a custom in-house dataset to address this gap. This dataset, sourced from Bulekha TV[7], represents the variety of Punjabi spoken in Pakistan. As the available data in the broadcast domain was limited, the recordings primarily comprised vlogs. These recordings were first converted to .wav format with specified properties: mono channel, 256 kbps bitrate, and 16 kHz sampling rate.

The dataset covers diverse topics relevant to the Punjabi-speaking audience. Annotation was carried out using XTrans (Glenn et al., 2009) in the Punjabi Shahmukhi script by trained annotators. For this study, carefully considered 15-hour dataset was used for few-shot learning, for ensuring balanced finetuning across all datasets to maintain consistency in performance evaluations.
For evaluation, 4.2 hours of data sourced from 52 speakers was utilized.

2.3 Urdu Dataset

Two datasets were used for the Urdu language: the Urdu Broadcast and Urdu Telephonic datasets, with detailed descriptions provided in the following subsections.

2.3.1 Urdu Broadcast Dataset

The Urdu Broadcast Dataset (Khan et al., 2021) contains approximately 800 hours of spoken Urdu from various broadcast platforms like Radio, YouTube, and TV. The dataset covers genres such as news, health, entertainment, and political discussions, capturing dialectal and phonetic variability. For this study, a thoughtfully chosen 15-hour dataset from 131 speakers was used for few-shot learning, while 4.3 hours from 45 speakers were allocated for evaluation, covering a wide range of regional accents and demographics.

2.3.2 Urdu Telephonic Dataset

The Urdu Telephonic Dataset consists of 111.5 hours of read speech, balanced by gender and representing various districts of Pakistan. The dataset, recorded via laptop and telephone, captures conversational speech patterns typical in telephonic interactions. For this study, a carefully curated 15-hour dataset from 179 speakers was used for few-shot learning, while 10.2 hours from 60 speakers were set aside for evaluation, representing a variety of accents and age groups.

The Table 2 provides a breakdown of the datasets, including the fine-tuning and evaluation splits, as well as the total number of utterances for each dataset.

2.4 Pre-processing

Following pre-processing steps were implemented: (a) All audio files were converted to mono format with a sample rate of 16 kHz; (b) selected a subset of 15 hours from each dataset for few-shot fine-tuning; (c) ensured the audio segments were accurately aligned with their corresponding transcriptions; and (d) removed any unnecessary punctuation and characters from the transcriptions to maintain consistency. Additionally, the 15-hour subset was divided into 1-hour, 5-hour, 10-hour, and 15-hour splits for the purpose of few-shot experimentation.

3 Experiment

The evaluation consists of two phases: zero-shot evaluation and few-shot fine-tuning. In the zero-shot phase, Whisper-small, Whisper-medium, and Whisper-large are evaluated on various datasets (as detailed in Section 2). For few-shot fine-tuning,

[1] https://catalogue.elra.info/en-us/repository/browse/ELRA-W0092/
[2] https://www.youtube.com/@VOAPashto
[3] https://pa.azadiradio.com/
[4] https://www.voadeewanews.com/live/audio/49
[5] https://www.mashaalradio.com/
[6] https://www.shamshadtv.tv/
[7] https://www.youtube.com/c/BhulekhaTv

Language	Dataset	Small WER	Medium WER	Large WER
Pashto	Broadcast	98.43	99.04	85.60
Punjabi	Broadcast	86.83	86.04	54.73
Urdu	Broadcast	42.57	35.57	27.97
	Telephonic	70.09	62.12	46.64

Table 1: Zero-shot %WER for Whisper models (Small, Medium, Large) on Pashto, Punjabi and Urdu datasets

Language	Dataset	Fine-tuning Duration	Evaluation Duration	Fine-tuning Utterances	Evaluation Utterances
Pashto	Broadcast	15 h	4.8 h	7746	2226
Punjabi	Broadcast	15 h	4.2 h	13110	2361
Urdu	Broadcast	15 h	4.3 h	8206	2633
	Telephonic	15 h	10.2 h	14066	6358

Table 2: Fine-tuning and Evaluation Data Breakdown for Whisper on Pashto, Punjabi, and Urdu

Language	Dataset	1hr WER	5hrs WER	10hrs WER	15hrs WER
Pashto	Broadcast	53.08	40.33	36.38	34.10
Punjabi	Broadcast	45.18	41.57	41.80	38.01
Urdu	Broadcast	33.14	27.26	23.43	22.28
	Telephonic	74.16	66.40	63.42	62.01

Table 3: Fine-tuned %WER for Whisper Small on Pashto, Punjabi, and Urdu datasets

Whisper-small is selected due to hardware constraints, with 15 hours of labeled data from each dataset, split into 1-hour, 5-hour, 10-hour, and 15-hour subsets to analyze the effect of dataset size. The zero-shot performance of Whisper-large is compared with the few-shot fine-tuned Whisper-small, focusing on the reduction in WER between the models.

3.1 Experimental Setup

The experiments are conducted on a system with two NVIDIA RTX 3060 GPUs[8], each with 12 GB of VRAM. To manage memory constraints, gradient accumulation is employed during fine-tuning. The AdamW optimizer is used with a learning rate of 1e-5 and a warmup period of 500 steps for stability. Fine-tuning is performed for a maximum of 100 epochs, with early stopping after 3 epochs to prevent overfitting.

3.2 N-shot Learning

Whisper's performance was evaluated in zero-shot and few-shot settings across Pashto, Punjabi, and Urdu. Zero-shot results set a baseline, while few-shot fine-tuning demonstrates how WER reduces with increasing data, offering insights into the model's real-world potential. **Zero-shot Learning:** In all zero-shot evaluations, the target language for transcription was explicitly specified by passing its corresponding language code as a parameter to the model. Whisper transcribes Pashto with inconsistent script usage, occasionally switching between the script conventions used for Northern Pashto dialects and Southern Pashto dialects. This variability reflects regional differences in orthographic practices, which led to inconsistencies in the transcription output. Similarly, for the Punjabi dataset, Whisper defaulted to Gurmukhi script, despite the widespread use of Shahmukhi by 94.4 million users (Ahmad et al., 2020), leaving the Shahmukhi script underrepresented in the transcription process.

Few-shot Learning: In the few-shot phase, Whisper-small was fine-tuned on datasets in incremental batches of 1 hour, 5 hours, 10 hours, and a maximum of 15 hours. The 15-hour limit was maintained for all languages, as the Punjabi few-shot learning dataset only consisted of 15 hours of data. Each step of fine-tuning allowed the model to progressively refine its transcription accuracy, capturing the nuances of scripts, and accents as explained in detail in the later section.

[8] https://www.nvidia.com/en-us/geforce/graphics-cards/30-series/rtx-3060-3060ti/

4 Results

This section presents the performance of Whisper on Pashto, Punjabi, and Urdu, emphasizing the impact of few-shot fine-tuning on transcription accuracy.

Zero-shot Results: All the outputs were post-processed to remove any punctuations marks. For Pashto, transcription consistency was ensured by converting the script conventions used for Northern dialect into standard Afghan Pashto using GPT-prompt to compute the WER. Whisper Large achieved a WER of 85.60, reflecting challenges with script variations, while Whisper Medium and Small recorded WERs above 90. Despite explicitly specifying the target language, Whisper Small and Medium exhibited frequent language switching during transcription. For Punjabi, post-processing was performed to convert Gurmukhi to Shahmukhi script using a GPT-prompt, with Whisper Large recording a WER of 54.73, outperforming Whisper Small's WER of 86.83. For Urdu, Whisper Large excelled in the Broadcast dataset with a WER of 27.97, surpassing Whisper Small's 42.57. On the Telephonic dataset, Whisper Large achieved a WER of 46.64, significantly outperforming Whisper Small, which had a WER of 70.09. Further details on these performance discrepancies regarding Pashto, Punjabi and Telephonic Urdu Datsets are provided in Appendix A. Despite its overall superior performance, Whisper Large struggled with regional accents, necessitating further adaptation. The results of zero-shot evaluation are presented in Table 1.

Few-shot Learning Results: Fine-tuning Whisper Small with varying durations resulted in significant WER reductions. For Pashto, WER decreased from 53.08 to 34.10 after 15 hours of fine-tuning, hence improving transcription in standard Afghan Pashto demonstrating the impact of domain-specific data. In Punjabi, fine-tuning reduced WER from 45.18 to 38.01, enabling transcription in Shahmukhi script, which was previously rendered in Gurmukhi. For Urdu, fine-tuning yielded substantial improvements, lowering WER from 33.14 to 22.28 for Broadcast and from 74.16 to 62.01 for Telephonic, indicating better adaptation to formal broadcast speech. The results are shown in Table 3.

An interesting observation is that fine-tuning Whisper Small significantly narrows the gap with Whisper Large in zero-shot performance. For Pashto, WER dropped from 53.08 to 34.10, surpassing Whisper Large's 85.60. In Punjabi, WER decreased from 45.18 to 38.01, outperforming Whisper Large's 54.73. For Urdu Broadcast, WER improved from 33.14 to 22.28, exceeding Whisper Large's 27.97. However, for Urdu Telephonic, WER dropped from 74.16 to 62.01, but Whisper Large's 46.64 still outperformed the fine-tuned model. These results demonstrate that fine-tuning Whisper Small with domain-specific data leads to substantial improvements across languages and datasets, significantly reducing the performance gap with Whisper Large.

5 Conclusion

This research highlights the effectiveness of fine-tuning Whisper models for low-resource languages like Pashto, Punjabi, and Urdu. While Whisper Large excelled in zero-shot evaluation, fine-tuning Whisper Small with domain-specific data led to substantial improvements in transcription accuracy. The significant reductions in WER across these languages demonstrate the power of fine-tuning to optimize performance and adapt Whisper to the unique linguistic characteristics of low-resource settings.

6 Limitation

This study has several limitations. Due to GPU resource constraints, fine-tuning was limited to Whisper Small, restricting the model's full potential. With access to more computational resources, fine-tuning Whisper Medium or Large could have enhanced performance across a wider range of datasets. Furthermore, the evaluation datasets for both Pashto and Punjabi were limited to a single dialect, which may not fully capture the linguistic diversity present within these languages. Additionally, Whisper would benefit from more diverse fine-tuning data, particularly for low-resource dialects, to improve generalization and achieve better results.

References

Muhammad Tayyab Ahmad, Muhammad Kamran Malik, Khurram Shahzad, Faisal Aslam, Asif Iqbal, Zubair Nawaz, and Faisal Bukhari. 2020. Named entity recognition and classification for punjabi shahmukhi. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(4):1–13.

Samee Arif, Aamina Jamal Khan, Mustafa Abbas, Agha Ali Raza, and Awais Athar. 2024. Wer we stand: Benchmarking urdu asr models. *arXiv preprint arXiv:2409.11252*.

Andrea Do, Oscar Brown, Zhengjie Wang, Nikhil Mathew, Zixin Liu, Jawwad Ahmed, and Cheng Yu. 2023a. Using fine-tuning and min lookahead beam search to improve whisper. *arXiv preprint arXiv:2309.10299*.

X. Do et al. 2023b. Enhancing asr performance with low-rank adaptation. *Journal of Speech Technology*.

Meghan Lammie Glenn, Stephanie M Strassel, and Haejoong Lee. 2009. Xtrans: A speech annotation and transcription tool. In *Tenth Annual Conference of the International Speech Communication Association*.

Erbaz Khan, Sahar Rauf, Farah Adeeba, and Sarmad Hussain. 2021. A multi-genre urdu broadcast speech recognition system. In *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 25–30. IEEE.

Panteleimon Krasadakis, Evangelos Sakkopoulos, and Vassilios S Verykios. 2024. A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages. *Electronics*, 13(3):648.

Yunpeng Liu and Dan Qu. 2024. Parameter-efficient fine-tuning of whisper for low-resource speech recognition. In *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, pages 1522–1525. IEEE.

Gaurav Maheshwari, Dmitry Ivanov, Théo Johannet, and Kevin El Haddad. 2024. Asr benchmarking: Need for a more representative conversational dataset. *arXiv preprint arXiv:2409.12042*.

Riefkyanov Surya Adia Pratama and Agit Amrullah. 2024. Analysis of whisper automatic speech recognition performance on low resource language. *Jurnal Pilar Nusa Mandiri*, 20(1):1–8.

A. Radford, J. Wu, and R. Child. 2022. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*.

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.

Munaza Sher, Nasir Ahmad, and Madiha Sher. 2024. Towards end-to-end speech recognition system for pashto language using transformer model. *IJIST*, 6 (1):115–131.

Bashar Talafha, Abdul Waheed, and Muhammad Abdul-Mageed. 2023. N-shot benchmarking of whisper on diverse arabic speech recognition. *arXiv preprint arXiv:2306.02902*.

Suhas Waghmare, Chirag Brahme, Siddhi Panchal, Numaan Sayed, and Mohit Goud. 2023. Comparative analysis of state-of-the-art speech recognition models for low-resource marathi language. *International Journal of Innovative Science and Research Technology (IJISRT)*, pages 1544–1545.

Chih-Kai Yang, Kuan-Po Huang, Ke-Han Lu, Chun-Yi Kuan, Chi-Yuan Hsiao, and Hung-yi Lee. 2024. Investigating zero-shot generalizability on mandarin-english code-switched asr and speech-to-text translation of recent foundation models with self-supervision and weak supervision. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 540–544. IEEE.

A Discussion on Errors in Zero-Shot Evaluation

In this section, we analyze the performance of different Whisper model variants (Small, Medium, and Large) on Pashto, Punjabi, and Urdu datasets in a zero-shot setting. The primary focus is on evaluating the transcription

errors observed in each language and understanding the limitations of the models in detail. Table 4 provides a comparative overview of the outputs from each model. Whisper Small and Medium produced unintelligible outputs, mixing scripts like Khmer and Telugu (e.g., " "), and generating gibberish, especially in Pashto and Punjabi (e.g., "livel, ündarvs"). In Urdu, they misinterpreted numerals and proper nouns (e.g., "gnignnaM "). Whisper Large performed better but still missed key contextual phrases in Pashto (e.g., missing " " – "Accept my greetings"), had subtle phonetic errors in Punjabi (e.g., misinterpreting " " as "old master"), and struggled with numerals in Urdu (e.g., " " instead of "twenty-seventh"). Overall, Whisper Large showed improvement but still faced significant limitations, indicating the need for further pretraining to improve zero-shot performance on these languages.

Language	Reference	Small	Medium	Large
Pashto	د ازادی رادیوي مجله زما اسد الله غضنفر سلامونه ومنئ قدرمنو اوریدونکو *("Azadi Radio Magazine, this is Asadullah Ghazanfar. Accept my greetings, dear listeners.")*	ila Di Vo Soga al H expected Emread this اګلا	livel, ündarvs ઽઽઽ snail iziert <lrul> säga selamun hpanaatge el ani beit cor generation తలం వలాయ భూసిల లలో	ده ازادی راډیوی مجلّه او پا دی بارا که بحث لرو *("This is the Azadi Radio Magazine, and we have a discussion on this topic.")*
Punjabi	ترے جنوبی ایشیائی نیں اے انہاں تناں دا پرانا آقا برطانیہ مقابلے ے ھے *("This is not your South Asia; here, the old master of all three, Britain, is in competition.")*	వసినినిందిని	Vermikha sharwa sa arpaki ya akarpesha	ਤਰੇ ਜੁਨੂਬੀ ਏਸ਼ਿਆਈ ਨੇ ਉਤੇ ਇਨਾ ਤੀਨਾ ਦਾ ਪੁਰਾਨਾ ਆਕਾ ਬਰਤਾਨੀਆ ਮੁਕਾਬਲੇ ਤੇ ਹੇ *("The South Asian stars are facing their old boss Britain.")*
Urdu	کیا ٹوینٹی سیونتھ اپریل ٹوینٹی فرسٹ تک ڈیم بن پائے گا *("Will the dam be completed by the twenty-seventh of April or by the twenty-first?")*	atra 둫	سکوہ Mannging کیا *("Will Mannging skouh?")*	کیا دونٹی سیونی ڈیٹل ڈی ساؤزن ڈی فورس ڈینڈ ڈینڈ جائے گا *("Will Donte Sioni detal de sauzen de force dand dand go?")*

Table 4: Whisper Model Responses Comparison for Different Languages (Pashto, Punjabi, and Urdu) in Zero-Shot Evaluation