WER We Stand: Benchmarking Urdu ASR Models

Samee Arif[1], Aamina Jamal Khan[1*], Mustafa Abbas[1*],
Agha Ali Raza[1], Awais Athar[2],
{samee.arif, 25100162, 25100079, agha.ali.raza}@lums.edu.pk,
awais@ebi.ac.uk
[1]Lahore University of Management Sciences,
[2]EMBL European Bioinformatics Institute

Abstract

This paper presents a comprehensive evaluation of Urdu Automatic Speech Recognition (ASR) models. We analyze the performance of three ASR model families: `Whisper`, `MMS`, and `Seamless-M4T` using Word Error Rate (WER), along with a detailed examination of the most frequent wrong words and error types including insertions, deletions, and substitutions. Our analysis is conducted using two types of datasets, read speech and conversational speech. Notably, we present the first conversational speech dataset designed for benchmarking Urdu ASR models. We find that `seamless-large` outperforms other ASR models on the read speech dataset, while `whisper-large` performs best on the conversational speech dataset. Furthermore, this evaluation highlights the complexities of assessing ASR models for low-resource languages like Urdu using quantitative metrics alone and emphasizes the need for a robust Urdu text normalization system. Our findings contribute valuable insights for developing robust ASR systems for low-resource languages like Urdu.

1 Introduction

Automatic Speech Recognition (ASR) systems have become integral to modern technology, enabling numerous applications. Use cases include virtual assistants (Adline Freeda et al., 2024) (Subhash et al., 2020), smart homes (Chen and Zhang, 2019) (Caranica et al., 2017), medical assistance (Maier et al., 2010) (Johnson et al., 2014), telecommunications (Rabiner, 1997), and more. The ability to convert spoken language into text has revolutionized human-computer interaction, making technology more accessible and user-friendly.

ASR systems have made substantial progress in recent years, driven by advances in deep learning and the availability of large-scale datasets. How-

ever, the majority of this progress has been concentrated on resource-rich languages, leaving low-resource languages like Urdu with significant gaps in accuracy and reliability. Urdu, with over 70 million native speakers, is characterized by its rich phonetic diversity, complex morphological structure, and variety of regional dialects, all of which pose additional challenges for ASR systems. These challenges are further amplified in conversational contexts by informal speech patterns, code-switching (Khan et al., 2023), and spontaneous speech disfluencies. The availability of annotated datasets is also limited compared to high-resource languages, which hinders the training and evaluation of ASR models.

Addressing these challenges is crucial for making ASR technology accessible and effective for Urdu speakers, particularly in scenarios involving both read and conversational speech. This paper focuses on evaluating and improving the performance of state-of-the-art ASR models and post-processing techniques for Urdu. Specifically, we examine three prominent ASR model families—`Whisper` (Radford et al., 2022), `MMS` (Pratap et al., 2023), and `Seamless-M4T` (Communication, 2023a) (Communication, 2023b)—each of which has demonstrated strong performance across various languages.

In this paper, we introduce the first conversational speech dataset specifically designed for evaluating Urdu ASR models. We release all our fine-tuned models, datasets, evaluation scripts, and outputs to the community to foster further research in Urdu ASR tasks. By making these resources publicly available, we aim to bridge the gap in ASR technology for low-resource languages like Urdu, thereby contributing to more inclusive and effective speech recognition solutions.

Our contributions can be summarized as follows:

1. We present the first conversational speech dataset for benchmarking Urdu ASR models.

---

*These authors contributed equally to this work.

2. We fine-tune ASR models from the `Whisper`, `MMS`, and `Seamless-M4T` families on Urdu, using both read speech and conversational speech datasets.

3. We conduct a comprehensive quantitative and qualitative analysis of non-fine-tuned and fine-tuned ASR models. We also highlight the complexity of evaluating Urdu ASR models using quantitative metrics alone, underscore the need for a robust text normalization system to address variations in word forms and improve overall model accuracy.

The code, and model outputs are publicly available on GitHub[1].

2 Related Work

2.1 Modern ASR Models

Modern ASR models, include `Wav2Vec2` (Baevski et al., 2020), `Whisper`, `MMS` and `Seamless-M4T`. `Wav2Vec2`, introduced by Facebook AI, revolutionizes ASR by learning powerful speech representations from raw audio, enabling it to perform exceptionally well even with limited labeled data. `Whisper` by OpenAI, leverages large-scale datasets and transformers to achieve state-of-the-art accuracy across multiple languages. Similarly, Meta's `MMS` and `Seamless-M4T` models push the boundaries of multilingual and cross-modal ASR by incorporating vast amounts of diverse linguistic data. These advancements have significantly improved the ability of ASR systems to handle continuous, natural speech across a wide variety of languages and contexts.

2.2 Low-Resource ASR

The ASRoIL survey (Singh et al., 2020) provides valuable insights for Indian languages into the challenges of variability in speech signals and the availability of corpora, detailing feature extraction techniques like MFCC, LPCC, and PLP, as well as various modeling and classification techniques. Unnibhavi and Jangamshetti (2016)'s work offers an overview of ASR systems for South Indian languages, explaining methodologies and feature extraction methods by giving digests of seven papers. Javed et al. (2022) introduce the IndicSU-PERB benchmark, featuring 1,684 hours of labeled speech data and establish benchmarks for

six speech tasks. They train and evaluate different self-supervised models and demonstrate that language-specific fine-tuned models outperform baseline methods. Dhouib et al. (2022)'s work systematically reviewed Arabic ASR studies from 2011 to 2021, covering feature extraction and classification techniques along with various aspects, such as the types of Arabic language supported, performance metrics, and existing research gaps. Abdelhamid et al. (2020) focus on end-to-end deep learning frameworks, which integrate and simultaneously train the language model, pronunciation, and acoustic components, thus simplifying the pipeline. Kadyan et al. (2019) presents a comparative study of Deep Neural Network (DNN) based Punjabi-ASR systems demonstrating that DNN-HMM hybrid models outperform traditional GMM-HMM architectures.

Low-resource languages have demonstrated significant advancements in multilingual settings, where models benefit from unexpected but advantageous learning capabilities. Toshniwal et al. (2018) found that multilingual models exhibit cross-lingual transfer learning, which enhances the accuracy of low-resource languages by leveraging similarities and commonalities from high-resource languages. Tüske et al. (2013) further analyzed cross-lingual transfer effects, finding that multilingual models fine-tuned on low-resource languages outperformed monolingual models, demonstrating the effectiveness of cross-lingual transfer. Overall, their work shows that low-resource languages like Urdu are more sensitive to model architecture due to data scarcity, and multilingual models can be trained effectively for both low-resource multilingual and monolingual tasks.

2.3 Urdu ASR

Sharif et al. (2024) provide an amalgamation of overviews of relevant studies on Urdu ASR, highlighting current trends, technological advancements, and future research directions, without directly comparing results on a common dataset. Farooq et al. (2019) develop an Urdu LVCSR system using 300 hours of speech data and explores various acoustic modeling techniques, achieving a WER of 13.50%, though the models and data are not publicly available. Ashraf et al. (2010) developed a small-sized GMM using Word List Grammar by processing individual words and combining it with a Knowledge Base storing audio representations, pronunciations, and probabilities, achieving

---

[1] https://github.com/sameearif/WER-We-Stand

a Word Error Rate (WER) of 5.33 on seen speakers and 10.66 on unseen speakers. Asadullah et al. (2016) explored the impact of vocabulary size on HMM performance for Urdu ASR, finding an insignificant WER increase of 0.5% between 100 and 250-word vocabularies. Naeem et al. (2020) used a GMM combined with a graphene-to-phone LSTM CMUsphinx and KENLM to calculate and store n-gram probabilities, achieving a 9.64% WER, significantly improving accuracy. Khan et al. (2023) combined HMM and CNN-TDNN with LF-MMI, achieving a 13% WER.

3 Models and Datasets

3.1 ASR and Post-Processing Models

In this paper we evaluate 9 ASR models given in Table 1.

Family	Model
Whisper	Tiny, Base, Small, Medium, Large-v3
MMS	300 Million, 1 Billion
Seamless-M4T	Medium, v2-Large

Table 1: ASR model evaluated in this paper.

3.2 ASR Datasets

We evaluate the ASR models on two types of datasets: (1) Read Speech, and (2) Conversational Speech. We developed and present the conversational speech dataset in this paper.

3.2.1 Read Speech

We use the ARL Urdu Speech Database[2] which has 159,996 audios. The distribution of speaker dialects in the corpus is given in Table 2.

For the dataset creation each speaker is presented with 400 prompts to read: sentences, place names, and person names. Two microphones set at different distances to the speaker are used for the recordings. Punctuation are omitted and numbers were written out in full.

We also evaluate the models using the bona fide audio files from the CSaLT Deepfake dataset (Munir et al., 2024), which includes 6 female and 11 male speakers, with a total of 3,398 audio files amounting to 42.9 hours of recordings.

---

[2] https://catalog.ldc.upenn.edu/LDC2007S03

Accent	Number of Speakers
South Sindh	29
North Sindh	30
South Punjab	27
North Punjab	29
Capital Area	29
North West Regions	30
Baluchistan	26

Table 2: Distribution of Speakers Across Different Accents in the Dataset

3.2.2 Conversational Speech

We present the first Urdu conversational speech dataset, which consists of 471 audio recordings (1.3 hours). These audios were recorded through calls over the internet. This is to mirror actual conversational environments and meetings, making the data more relevant and practical for real-world applications.

The dataset features 4 female and 6 male speakers who used the microphones they had on hand to replicate real-life audio quality. To ensure smooth and natural conversations, the participants, all native Urdu speakers and computer science students, were asked to form groups and pairs with people they felt comfortable with. The recordings were done in various group sizes. Four sessions involved groups of 2, One session involved a group of 3, and another session involved a group of 4.

Participants were also encouraged to choose topics they were comfortable discussing, which resulted in a diverse range of discussions. The final topics included Pakistan Independence Day, Group Projects, Ramadan and Eid, Neighbors Discussing Load-shedding and Prices, Health Issues, and Gossips.

For transcription, the process was carried out in three passes to ensure accuracy and consistency. The first pass transcription was completed by the original recorders themselves. This was followed by a second pass, where two research interns, who were not involved in the recording, refined the transcriptions and split the audio into smaller chunks. The final pass was done by two of the authors of this paper, ensuring the highest level of accuracy in the transcriptions.

4 Experimental Design

We fine-tune the ASR models in Table 1 using a combined dataset of Mozilla's Common Voice (Ardila et al., 2020) and Google's Fleurs (Conneau et al., 2022), which together provide 16,156 audio samples. Given the limited amount of training data available, we opted for a 90-10 split, with 90% of the data used for training and the remaining 10% for validation. We evaluate both the non-fine-tuned and the fine-tuned models on ARL and CSaLT dataset for read speech and on our dataset for conversational speech. `mms-300m` base model is not fine-tuned for downstream task of speech recognition so we only evaluate our fine-tuned version. Before conducting the evaluations, we normalize both the ground truth and the model predictions. This involves removing punctuation and disfluencies. This preprocessing step ensures that the evaluations focus on meaningful transcription accuracy rather than being skewed by minor formatting inconsistencies or speech disfluencies and punctuation marks. We noticed that Whisper and MMS models have built-in disfluency filtering and perform this task effectively. However, Seamless models output disfluencies marked with a # (e.g., #um), which we removed using regex during preprocessing.

5 Results and Discussion

Table 3 shows the WER of the ASR models, fine-tuned and non-fine-tuned, on read speech datasets—ARL and CSaLT—and our conversational speech dataset.

5.1 Read Speech

Read speech typically consists of well-articulated sentences with fewer disfluencies, making it a less complex task for ASR models compared to conversational speech. However, challenges such as dialectal variations and pronunciation differences still impact transcription accuracy.

5.1.1 Quantitative Analysis

In both the ARL and CSaLT datasets, the evaluated ASR models reveal distinct patterns in performance based on their architecture and size. Smaller models, such as `whisper-tiny` and `whisper-base`, suffer from significant hallucination issues in their base versions, as indicated by the high WERs—116.92 and 96.57 on ARL for `whisper-tiny` and `whisper-base`, respectively.

Similarly, on CSaLT, the WER for `whisper-tiny` is 96.57, indicating that these models struggle to handle Urdu without fine-tuning. This issue is depicted in Figure 1, where the non-fine-tuned `whisper-tiny` generates incorrect transcriptions. However, fine-tuning substantially reduces these error rates, as seen with `whisper-tiny`'s drop to 45.59 on ARL and 42.12 on CSaLT, and `whisper-base`'s improvement to 39.84 and 38.86 on the respective datasets. Despite these improvements, the smaller models remain limited due to their relatively small parameter sizes—34M for `whisper-tiny` and 74M for `whisper-base`.

یک ایک ایک ایک ایک ایک ایک ایک ایک ایک ایک ایک ایک ایک ایک ایک ایک ایک
ایک ایک ایک ایک ایک ایک ایک ایک ایک ایک ایک ایک ایک ایک ایک ایک ایک...

Aik aik aik aik aik aik aik aik aik aik aik aik aik aik aik aik aik aik aik aik aik aik aik aik aik aik aik…

ایسے خطے میں گھری اسکول سے دور پر ہائی مشکل ہے جہاں گئی گھروں میں
انڈنٹ کی سہولت نہیں اور سمارٹ فون تک رسائی بھی محدود ہے

Aise khitte mein ghiri school se door par high mushkil hai jahan gayi gharon mein internet ki sahulat nahin aur smartphone tak rasai bhi mehdood hai.
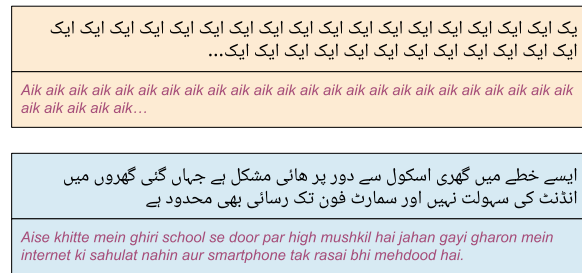
Figure 1: The yellow and blue boxes represent the output of the same audio on general and fine-tuned `whisper-tiny` models respectively

Moving to larger models, `whisper-small` (244M parameters) and `whisper-medium` (769M parameters) demonstrate a notable improvement over their smaller counterparts. For instance, `whisper-small` reduces its WER from 48.70 to 28.60 on ARL after fine-tuning, while `whisper-medium` improves from 37.04 to 25.38. The trend is consistent on CSaLT, where `whisper-small` drops from a WER of 41.10 to 27.39 and `whisper-medium` from 33.39 to 24.15. However, the largest model, `whisper-large` with 1.55B parameters, performs better, achieving a WER of 23.79 on ARL and 22.35 on CSaLT after fine-tuning. Although `whisper-large` shows impressive results, the `seamless-m4t` models outperform it. For instance, `seamless-medium` with 1.2B parameters achieves a WER of 30.06 pre-fine-tuning and 19.41 post-fine-tuning on ARL. Additionally, fine-tuned `seamless-large` consistently outperforms all other models, with a WER of 17.09 on ARL and 18.61 on CSaLT, showcasing the efficiency of the Seamless-M4T architecture.

In contrast, the `mms` models, despite their large parameter sizes, underperform. The `mms-300m` model struggles, with a WER of 51.48 on ARL and 47.73 on CSaLT datasets, whereas `mms-1b` achieves more competitive results after fine-tuning, reducing

	Read Speech (ARL)		Read Speech (CSaLT)		Conversational Speech	
Model	Base Model	Fine-tuned	Base Model	Fine-tuned	Base Model	Fine-tuned
whisper-tiny	116.92	45.59	96.57	42.12	163.18	59.99
whisper-base	71.53	39.84	57.77	38.86	163.52	48.61
whisper-small	48.70	28.60	41.10	27.39	55.67	32.92
whisper-medium	37.04	25.38	33.39	24.15	40.22	28.87
whisper-large	26.25	23.79	24.44	22.35	**18.30**	**17.86**
mms-300m	-	51.48	-	47.73	-	66.40
mms-1b	39.65	28.37	34.60	26.85	46.44	42.46
seamless-medium	30.06	19.41	24.18	20.59	22.33	20.01
seamless-large	**23.97**	**17.09**	**20.57**	**18.61**	29.99	18.75

Table 3: WER of ASR models on ARL and CSaLT dataset for Read Speech and on our dataset for Conversational Speech. For each category, bold represents the lower WER in the specified category.

its WER from 39.65 to 28.37 on ARL and from 34.60 to 26.85 on CSaLT. mms-300m has a lower WER on both datasets compared to whisper-tiny with 39M parameters. The fine-tuned mms-1b has comparable performance to whisper-small with 244M parameters. This indicates the MMS family is outperformed by much smaller models from other model families, highlighting that model size alone does not guarantee better performance, and other factors such as architecture and training data play a crucial role.

The heatmap in Figure 2 compares the most frequent errors (wrong words) on combined ARL and CSaLT datasets across five non-fine-tuned models: mms-1b, whisper-medium, whisper-large, seamless-medium, and seamless-large.
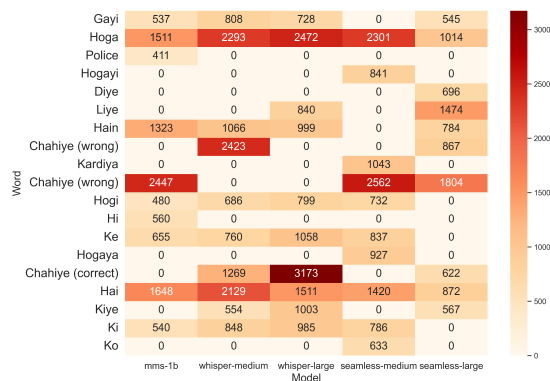

Figure 2: Comparison of wrong words across non-fine-tuned models.

Several words appear to stand out as high-frequency errors across all models. Notably, "Ho ga" and "Chahiye" are consistently problematic, with particularly high error counts in whisper-medium, whisper-large, and seamless-large. The word "Chahiye" appears with orthographic variations in Urdu—one with two "Yeh" and no "Hamza" character, one with one "Yeh" and a "Hamza," and one with with no "Yeh" and a "Hamza." The first variation is the most widely accepted correct form, though all are used interchangeably in practice. These errors can likely be attributed to the training data, which often includes user-generated content like internet text, where such variations are common.

"Chahiye" has an error frequency of 1,269 in whisper-medium and 3,173 in whisper-large, indicating that these models do not reproduce the specific spelling of this word as used in the dataset. mms-1b shows relatively fewer errors for many words in this specific analysis, but it has a higher WER for the ARL and CSaLT datasets. This indicates that, while fewer individual words may be highlighted as problematic here, the model struggles with a broader range of words overall, leading to a higher overall error rate. The word "Hoga" shows a high error rate across all models, as it is reproduced as "Ho ga" in the model transcriptions, but the test dataset contains it as a single word, "Hoga." This mismatch leads to consistent false positives, where the models' correct transcription is flagged as incorrect due to the way the word is represented in the test dataset. The Whisper family has the higher tendency to misrecognize the word "Hai" followed by MMS and then Seamless.

Figure 3 compares the most frequent errors on ARL and CSaLT datasets across the fine-tuned models. In the fine-tuned models, certain high-frequency errors observed in the non-fine-tuned models have been reduced, indicating that fine-tuning has helped address some transcription challenges. However in some cases the error rate has gone up. For instance, the error rate for the word "Chahiye" (expected spellings) has increased after

fine-tuning. We found 155 misspelled versions of this word in Common Voice and Fleurs dataset. For `seamless-medium`, the error has gone up from 0 to 1942. `whisper-large` still shows a high error with slight increase from 3173 to 3391. The word "Ho ga" also poses a challenge for fine-tuned models. There are 82 instances of "Ho ga" and 87 instances of "Hoga" in the training dataset. As a result, the false positives seen earlier are now turned into inconsistencies, where the models are confused between the two variations, leading to unpredictable outputs. A lattice of plausible respellings of the reference transcription, as done so by Karita et al., 2023 may help address this issue. By employing a lenient evaluation with orthographical variations in mind, these false errors could be overlooked.
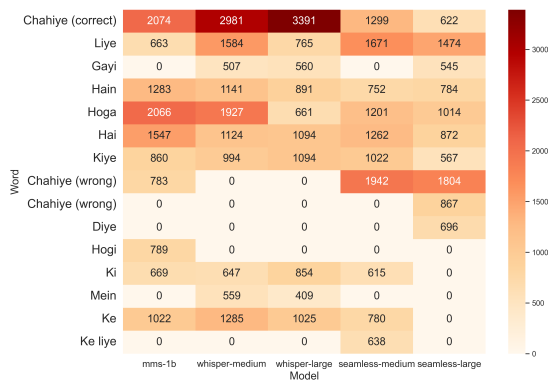
Figure 3: Comparison of wrong words across fine-tuned models.

The words like "Hi", "Ki" and "Ko" show improved performance in the fine-tuned models. While fine-tuning has led to notable improvements in transcription accuracy for several words, persistent challenges remain, particularly for high-frequency words and segmentation issues. This evaluation highlights the complexity of evaluating Urdu ASR models using quantitative metrics alone and underscores the need for a robust text normalization system to address variations in word forms and improve overall model accuracy.

5.1.2 Qualitative Analysis

Fine-tuning leads to substantial improvements across all models, the `seamless-large` emerges as the best performer across the ARL and CSaLT datasets. The performance of the `seamless-medium` and `seamless-large` models can be attributed to several key differences in training data and architecture. Seamless-M4T benefited from over 470,000 hours of multimodal speech and text data, sourced from the SeamlessAlign

dataset, which was specifically crafted to support a broad spectrum of languages. This extensive, diverse dataset, combined with the use of SONAR embeddings—designed to offer modality and language-agnostic representations—enhances Seamless-M4T's ability to generalize effectively across various languages and speech domains. The architecture supports more complex multimodal learning, contributing to its edge in performance, particularly in tasks involving low-resource languages like Urdu.

In contrast, `whisper-large` was trained on approximately 1 million hours of labeled data and 4 million hours of pseudo-labeled data, with a focus on automatic speech recognition (ASR) tasks. While Whisper is capable of handling numerous languages, its emphasis is primarily on transcription accuracy. Without the same multimodal and speech-to-speech training exposure found in Seamless-M4T, Whisper lacks the cross-modal adaptability that contributes to the superior results seen in Seamless-M4T's performance on read speech tasks.

Similarly, while the MMS models were pre-trained on 500,000 hours of speech data across 1,400 languages, their focus on Wav2Vec2-based self-supervised learning may explain their relative underperformance. MMS models are designed to excel at learning robust speech representations without labeled data, but they lack the comprehensive multimodal and multitask learning present in Seamless-M4T. This limits their performance on specific tasks like Urdu read speech, where cross-modal training and fine-tuning play a crucial role. Thus, despite extensive pretraining, MMS models are unable to match the fine-tuned, multimodal capabilities of Seamless-M4T in such applications.

5.2 Conversational Speech

The results for conversational speech highlight distinct performance patterns among ASR models when compared to read speech, mainly due to the inherent challenges posed by spontaneous, natural dialogue. Conversational speech often includes disfluencies, overlaps, and variations in speaker styles, making it more complex to transcribe accurately.

5.2.1 Quantitative Analysis

Most models show higher WERs in conversational speech, reflecting the increased difficulty of this task. Among the Whisper models, `whisper-large` stands out with the lowest WER of 18.30 in

its base version, slightly improving to 17.86 after fine-tuning. This positions it as the best-performing model for conversational speech. In contrast, smaller models like `whisper-tiny` and `whisper-base` struggle significantly. Their base versions exhibit extremely high WERs—163.18 and 163.52, respectively—due to hallucinations and transcription errors. Fine-tuning reduces these errors, bringing their WERs down to 59.99 and 48.61, but they remain far behind the larger models. Mid-range models, such as `whisper-small` (244M parameters) and `whisper-medium` (769M parameters), also improve with fine-tuning, achieving WERs of 32.92 and 28.87, respectively.

The MMS models, despite their large parameter sizes, underperform in conversational speech. The `mms-1b` model, shows some improvement after fine-tuning, reducing its WER from 46.44 to 42.46, but it still lags behind even the smaller Whisper models like `whisper-medium`. The `mms-300m` model performs worse, with a post-fine-tuning WER of 66.40, suggesting that the MMS models' pretraining data and objectives do not generalize well to conversational speech in low-resource languages such as Urdu.

The Seamless-M4T models also perform well on conversational speech, though their strengths lie primarily in read speech tasks. Notably, the smaller `seamless-medium` model initially outperforms its larger counterpart, `seamless-large`, in its base version, achieving a WER of 22.33 compared to 29.99 for the large model. This suggests that the medium model, despite having fewer parameters, handles the nuances of conversational speech better. However, after fine-tuning, `seamless-large` shows more significant improvement, reducing its WER to 18.75, while `seamless-medium` only improves to 20.01. This indicates that while the medium model may have an edge in its base state, the larger model benefits more from fine-tuning, ultimately surpassing the performance of the medium model.

Overall, the results for conversational speech show that larger Whisper models and Seamless-M4T are better equipped to handle the complexities of spontaneous dialogue. Whisper models, particularly `whisper-large`, exhibit a slight edge over Seamless-M4T, although the gap is narrower compared to their performance in read speech tasks. The MMS models, despite their large size, do not perform as well in this task, highlighting potential limitations in their training objectives. Fine-tuning

proves to be essential for improving performance across all models when dealing with conversational speech.

The analysis of the conversational dataset on non-fine-tuned models reveals significant variability in error patterns across the different ASR models. While all models show high substitution errors, there are notable differences in the number of deletions and insertions. Figure 4 gives the comparison of substitution errors and Figure 5 gives the comparison of deletion errors across models before and after fine-tuning. Substitutions dominate the total errors for all models, with counts ranging from 13,665 for `seamless-large` to 14,552 for `mms-1b`, indicating that non-fine-tuned models frequently mis-recognize words, leading to incorrect transcriptions. This is especially problematic for conversational datasets where context and word prediction are key factors for accurate transcription.
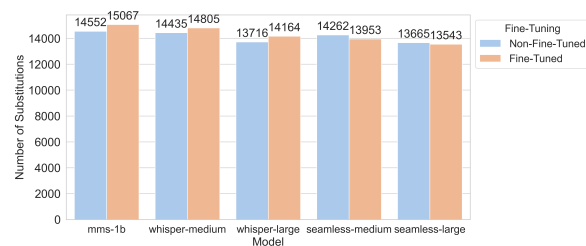

Figure 4: Comparison of substitution errors across models before and after fine-tuning.
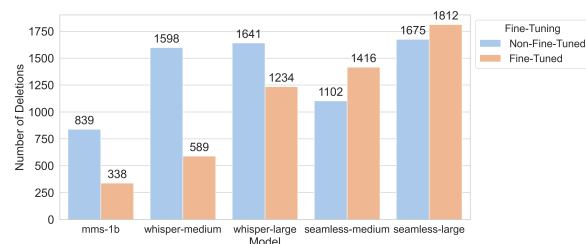

Figure 5: Comparison of deletion errors across models before and after fine-tuning.

Deletions also account for a substantial portion of errors, particularly for `whisper-large` (1,641) and `whisper-medium` (1,598). High deletion errors suggest that these models are often failing to transcribe words altogether, which can significantly distort the meaning of the conversation. Interestingly, `mms-1b` shows the lowest number of deletions (844), indicating that while it may substitute many words, it tends to capture more of the spoken content compared to other models. Insertions remain minimal across all models (non-fine-tuned

and fine-tuned), with each registering only a few insertion errors (1 to 4). This suggests that these models rarely add extra words that were not spoken.

After fine-tuning, the overall performance of the ASR models shows notable changes, though the error patterns remain similar in terms of overall trends. Substitution errors continue to account for the majority of the total errors. We can observe in Figure 4 and 5 that after fine-tuning the WER of Whisper models and MMS goes down because of reduction in deletion errors and the WER of Seamless models go down because of the reduction in substitution errors. Fine-tuning seems to reduce substitution errors for `seamless-large` to 13,543 and `seamless-medium` to 13,953 but it increased slightly for other models. Deletions error particularly, for `mms-1b` saw a drop from 844 to 338, and for `whisper-medium` dropped from 1,598 to 589. This indicates that fine-tuning helps the models transcribe more spoken content in their transcriptions.

5.2.2 Qualitative Analysis

A significant limitation across all ASR models is their inability to handle overlapping speech from multiple speakers. Figure 6 illustrates an example of ground truth for overlapping speech, where the audio from speaker 1 and speaker 2 overlaps. As shown in the figure, speaker 1's audio is dominant, causing ASR models to fail in accurately transcribing the speech of speaker 2.
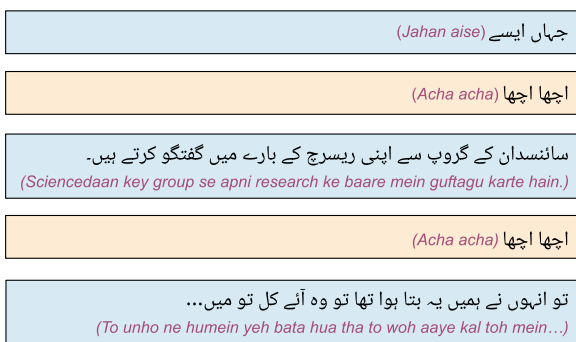

Figure 6: Example of multi-speaker audio. The blue color represents speaker 1 and yellow represents speaker 2.

The higher WER of base `seamless-large` can be partly attributed to two distinct factors: its handling of English words within the Urdu conversational dataset and its tendency to paraphrase certain words. Instead of transliterating English words, it often translates them, leading to errors. For exam-

ple, as shown in Figure 7, the word "Research" is translated to "Tehqeeq" rather than being transliterated. Additionally, the model tends to paraphrase words within the Urdu content, such as transcribing "Guftagu" as "Baat". These errors likely stem from the model's training on a multimodal dataset that emphasizes cross-language translation and broader contextual understanding, rather than strict transcription accuracy. Seamless-M4T's architecture, designed for tasks beyond ASR—such as speech-to-speech translation—encourages this behavior. The model prioritizes conveying meaning across languages, which leads to translations and paraphrasing when encountering multilingual inputs, as opposed to Whisper, which is primarily trained for transcription fidelity.
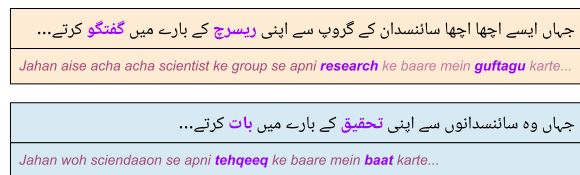

Figure 7: Translation and paraphrasing error in Seamless M4T. The yellow box represents the ground truth and the blue ones represents the prediction.

6 Conclusion

In this paper, we present a comprehensive evaluation of three ASR model families—Whisper, MMS, and Seamless-M4T—on Urdu read and conversational speech datasets. Our findings highlight the challenges associated with developing robust ASR systems for low-resource languages like Urdu, particularly when faced with spontaneous conversational speech, disfluencies, and code-switching. Among the models, `whisper-large` and `seamless-large` stand out, with `whisper-large` excelling in conversational contexts and `seamless-large` demonstrating strong performance in read speech tasks. Despite improvements from fine-tuning, challenges such as handling overlapping speech and distinguishing between similar phonetic patterns remain prevalent across models, indicating the need for further refinement. Additionally, our error analysis reveals the importance of text normalization and highlights the potential of multimodal approaches to improve ASR accuracy. This study contributes valuable insights into the capabilities and limitations of current ASR models for Urdu and underscores the importance of designing specialized datasets and

evaluation metrics. Our work will also be valuable for developers looking to build real-world applications, such as virtual assistants, voice-controlled devices, and transcription services.

7 Future Work

This study opens several avenues for future exploration in Urdu ASR. Addressing overlapping speech in conversational settings remains a critical challenge. Future work could focus on evaluating speaker diarization using our dataset to improve multi-speaker recognition. Enhancing the handling of code-switching between Urdu and English, especially for models like `seamless-large` that tend to translate rather than transliterate, could significantly boost transcription accuracy in real-world contexts where multilingualism is common. Exploring the potential of Large Language Models (LLMs), such as `GPT-4o` and `Llama-3.1`, for ASR output post-processing could further refine transcription quality by correcting errors and improving fluency. Finally, the integration of a lenient evaluation method which incorporates orthographical variations, or robust text normalization system tailored for Urdu could address variations in spelling and word forms, improving the consistency of transcriptions. Extending this work to cross-modal tasks, such as speech-to-speech translation or text generation, could provide valuable insights for building more advanced systems that handle complex, multimodal language tasks for low-resource languages like Urdu.

Acknowledgments

References

Abdelaziz Abdelhamid, Hamzah Alsayadi, Islam Hegazy, and Zaki Fayed. 2020. End-to-end arabic speech recognition: A review.

R. Adline Freeda, V. S. Krithikaa Venket, A. Anju, Gugan, Ragul, and Rakesh. 2024. Voice-based virtual assistant for windows using asr. In *ICT: Smart Systems and Technologies*, pages 277–284, Singapore. Springer Nature Singapore.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Asadullah, Arslan Shaukat, Hazrat Ali, and Usman Akram. 2016. Automatic urdu speech recognition using hidden markov model. In *2016 International Conference on Image, Vision and Computing (ICIVC)*, pages 135–139.

Javed Ashraf, Naveed Iqbal, Naveed Sarfraz Khattak, and Ather Mohsin Zaidi. 2010. Speaker independent urdu speech recognition using hmm. In *2010 The 7th International Conference on Informatics and Systems (INFOS)*, pages 1–5.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Alexandru Caranica, Horia Cucu, Corneliu Burileanu, François Portet, and Michel Vacher. 2017. Speech recognition results for voice-controlled assistive applications. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–8.

Hong Chen and Bo Zhang. 2019. Application of automatic speech recognition (asr) algorithm in smart home. *Journal of Physics: Conference Series*, 1237(2):022133.

Seamless Communication. 2023a. Seamless: Multilingual expressive and streaming speech translation. *Preprint*, arXiv:2312.05187.

Seamless Communication. 2023b. Seamlessm4t: Massively multilingual and multimodal machine translation. *Preprint*, arXiv:2308.11596.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *Preprint*, arXiv:2205.12446.

Amira Dhouib, Achraf Othman, Oussama El Ghoul, Mohamed Koutheair Khribi, and Aisha Al Sinani. 2022. Arabic automatic speech recognition: A systematic literature review. *Applied Sciences*, 12(17).

Muhammad Umar Farooq, Farah Adeeba, Sahar Rauf, and Sarmad Hussain. 2019. Improving large vocabulary urdu speech recognition system using deep neural networks. In *Interspeech 2019*, pages 2978–2982.

Tahir Javed, Kaushal Santosh Bhogale, Abhigyan Raman, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Indicsuperb: A speech processing universal performance benchmark for indian languages. *Preprint*, arXiv:2208.11761.

Maree Johnson, Samuel Lapkin, Vanessa Long, Paula Sanchez, Hanna Suominen, Jim Basilakis, and Linda Dawson. 2014. A systematic review of speech recognition technology in health care. *BMC medical informatics and decision making*, 14:94.

Virender Kadyan, Archana Mantri, Rajesh Aggarwal, and Amitoj Singh. 2019. A comparative study of deep neural network based punjabi-asr system. *International Journal of Speech Technology*, 22.

Shigeki Karita, Richard Sproat, and Haruko Ishikawa. 2023. Lenient evaluation of Japanese speech recognition: Modeling naturally occurring spelling inconsistency. In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, pages 61–70, Toronto, Canada. Association for Computational Linguistics.

Muhammad Danyal Khan, Raheem Ali, and Arshad Aziz. 2023. Code-switched urdu asr for noisy telephonic environment using data centric approach with hybrid hmm and cnn-tdnn. *Preprint*, arXiv:2307.12759.

Andreas Maier, Haderlein Tino, Florian Stelzle, Elmar Noeth, Emeka Nkenke, Rosanowski Frank, Schützenberger Anne, and Maria Schuster. 2010. Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010.

Sheza Munir, Wassay Sajjad, Mukeet Raza, Emaan Abbas, Abdul Hameed Azeemi, Ihsan Ayyub Qazi, and Agha Ali Raza. 2024. Deepfake defense: Constructing and evaluating a specialized Urdu deepfake audio dataset. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14470–14480, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Saad Naeem, Majid Iqbal, Muhammad Saqib, Muhammad Saad, Muhammad Soban Raza, Zaid Ali, Naveed Akhtar, Mirza Omer Beg, Waseem Shahzad, and Muhhamad Umair Arshad. 2020. Subspace gaussian mixture model for continuous urdu speech recognition using kaldi. In *2020 14th International Conference on Open Source Systems and Technologies (ICOSST)*, pages 1–7.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *Preprint*, arXiv:2305.13516.

L.R. Rabiner. 1997. Applications of speech recognition in the area of telecommunications. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 501–510.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Muhammad Sharif, Zeeshan Abbas, Jiangyan Yi, and Chenglin Liu. 2024. From statistical methods to pretrained models; a survey on automatic speech recognition for resource scarce urdu language. *Preprint*, arXiv:2411.14493.

Amitoj Singh, Virender Kadyan, Munish Kumar, and Nancy Bassan. 2020. Asroil: a comprehensive survey for automatic speech recognition of indian languages. *Artif. Intell. Rev.*, 53(5):3673–3704.

S Subhash, Prajwal N Srivatsa, S Siddesh, A Ullas, and B Santhosh. 2020. Artificial intelligence-based voice assistant. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, pages 593–596.

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.

Zoltán Tüske, Joel Pinto, Daniel Willett, and Ralf Schlüter. 2013. Investigation on cross- and multilingual mlp features under matched and mismatched acoustical conditions. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7349–7353.

Anand H. Unnibhavi and D. S. Jangamshetti. 2016. A survey of speech recognition on south indian languages. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pages 1122–1126.