A Benchmark Dataset and Comparative Evaluation of Phonemized and Romanized Urdu for Text-to-Speech

M Kaab Bin Shahid[†], **Muhammed Izharuddin**[‡]
[†]University of Stuttgart, Germany
[‡]Aligarh Muslim University, India
st200141@stud.uni-stuttgart.de, izharuddin@zhcet.ac.in

Abstract

Text-to-Speech (TTS) system for the Urdu language presents significant challenges, primarily due to the scarcity of high-quality datasets and an insufficient focus on modeling pronunciation. Urdu is spoken by 250 million people worldwide, but its research on computational linguistics remains underrepresented. In this paper, we introduce URDUTTS, a comprehensive and publicly available Urdu TTS dataset containing 89 hours of studio-quality speech, with accompanying transcriptions in three formats: Urdu Script, Phonemized Script, and Romanized Script. The dataset includes both mono-speaker and multi-speaker configurations. As Urdu relies heavily on phonetic features, accurate pronunciation is highly essential for the language. Therefore, we benchmark our dataset using VITS and GlowTTS models to compare the widely used Romanized script format with the Phonemized representation. To make the evaluation highly comprehensive, we combined both objective and subjective evaluation strategies. For objective evaluation, Mel-Cepstral Distortion (MCD with Plain, Dynamic Time-Warping, and Slope-Limitation variants), Signal-to-Noise Ratio (SNR), Word Error Rate (WER), and Character Error Rate (CER) were taken. Subjective evaluation was governed by Mean Opinion Score (MOS) ratings from 40 native speakers. Results show that using VITS and GlowTTS with Phonemized transcriptions performs significantly better than Romanized ones, with an improvement of 9.6% and 26.5% in MOS. The data and code are available at github.com/KAABSHAHID/URDUTTS.

Keywords: Low Resource Language, Text-to-Speech for Urdu, Benchmark Dataset Curation

1. Introduction

Urdu has an estimated 246 million speakers worldwide (Wikipedia contributors, 2025). Even with this large speaker community, computational linguistics (Amin et al., 2025) and speech technology research (Arif et al., 2025) for the language remains highly underrepresented. Particularly the domain of Text-to-Speech (TTS) for Urdu lags far behind compared to other major world languages. There have been various works on TTS systems for English, Chinese, and several regional South Asian languages, but research for Urdu has often been limited to small-scale with fine-tuned models built on low-quality datasets. The development of dedicated, large, and high-quality resources for Urdu has also been neglected.

Urdu TTS research also suffers from the bad treatment of pronunciation (Hussain, 2004; Usama et al., 2024). This language is phonetically rich and characterized by diverse classes of sounds (Niazi and Farhat, 2023): nasal consonants and nasalized vowels (e.g., /mã/), a complex plosive system with voiced/voiceless and aspirated/unaspirated contrasts (e.g., /k/ ک vs. /kʰ/ کھ), affricates (e.g., /tʃ/ چ, /dʒ/ ج), fricatives (e.g., /x/ خ, /ɣ/ غ), liquids (e.g., /l/, /r/), glides (e.g., /w/, /j/), retroflex and emphatic consonants (e.g., /ʈ/ ٹ, /ɽ/ ڑ), as well as a full set of oral and nasalized vowels. Additionally, Urdu inherits several pharyngeal and emphatic consonants from Arabic and Persian (e.g., /ħ/ ح, /ʕ/ ع), which further diversify its phonetic space. These sounds are the very backbone of the language's identity.

Unfortunately, previous research in Urdu TTS has often ignored these pronunciation complexities. Mainly the majority of the previous systems have directly relied on Romanized Urdu script which is represented using Latin characters. The problem with this is that the Romanized Urdu is highly ambiguous. For example, the Urdu word خواب (dream) can appear in Roman script as "khawab", "khwaab", or "khwab", depending on the writer. Hence, these representations fail to accurately capture aspirated sounds, vowel length, and pharyngeal contrasts. Similarly, nasalization and retroflexion are frequently lost, which leads to degraded pronunciation in synthesized speech. As a result, Romanized input fails to preserve the phonetic prosody that underpins Urdu's natural sound system.

1.1. Challenges in Urdu TTS

There are several challenges which makes Urdu a particularly difficult language for developing TTS systems:

Orthographic Complexity: Urdu uses a Perso-Arabic script (Nastaliq style), which omits short vowels in writing. For instance, the written form پر can be read in several ways: "par" (feather), "par"

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10982–10993
11–16 May 2026.

(but), or “pur” (full of, as in پُر خوشی “full of happiness”). All of these forms share the same written representation but they differ in pronunciation and meaning (Bashir, 2011). Hence, grapheme-to-phoneme (G2P) conversion becomes complicated as the correct reading generally depends on linguistic context.

Romanized Urdu Ambiguity: Romanized Urdu is widely used in informal digital communication, but it lacks standardization (Ahmed, 2009). The same word can have multiple Roman spellings, such as the Urdu word صاحب (master) can have various Roman spellings like “sahib” vs. “saheb” vs. “saahib”. Furthermore, many sounds have no direct Latin equivalent (e.g., retroflex /ʈ/, pharyngeal /ʕ/). Due to this lossy input in TTS models are inevitable.

Scarcity of Dataset: Existing Urdu speech datasets are extremely limited in size and quality. Earlier data were generally collected for unit-selection or HMM-based systems (Adeeba et al., 2016), which typically contain only a few hours of speech. Very limited works have used a good quality of dataset, but they are not publicly available (Naseem et al., 2025). Therefore, there is a scarcity of datasets that are large and simultaneously high quality in the public domain.

Pronunciation Modeling: Urdu has a rich phonetic inventory which requires (Niazi and Farhat, 2023) accurate handling of aspiration, retroflexion, nasalization, and pharyngeal. Previous works resulted in synthetic speech that is unnatural and has low pronunciation accuracy due to the less emphasis on these phonetic features.

Evaluation Limitations: Previous Urdu TTS studies have been evaluated either with small-scale intelligibility tests or by transferring models trained in other languages (Jamal et al., 2022). Comprehensive evaluations combining objective metrics (MCD, WER, CER, SNR) with subjective listening tests have rarely been performed (Naseem et al., 2025).

2. Prior Work

Urdu TTS has seen research going from traditional statistical-based to recent neural systems. Broadly, prior work can be grouped into four categories: (i) early statistical models, (ii) neural architectures with scratch training or fine-tuning, (iii) efforts around dataset development, and (iv) pronunciation modeling via script representation (Roman vs. Phoneme-based).

2.1. Traditional Approaches

Early Urdu TTS systems relied on unit-selection and Hidden Markov Model (HMM)-based synthesis. Habib (2014) adapted the HTS toolkit for Urdu by generating full-context labels and training HMM-based voices from scratch on limited data. Adeeba et al. (2016) compared unit-selection and HMM approaches using a 10-hour manually annotated studio data and reported 97.76% ASR accuracy on training data. Their results were intelligible but produced speech with robotic prosody and limited naturalness which reflects the well-known drawbacks of HMM-based synthesis (Tokuda et al., 2013).

2.2. Neural End-to-End Models

The advent of deep learning shifted Urdu TTS research toward neural architectures such as Tacotron (Wang et al., 2017), Tacotron 2 (Shen et al., 2018), and multi-speaker generative models. Saba et al. (2022) trained a Tacotron 2 + WaveGlow (Prenger et al., 2019) model from scratch on a 70-minute Urdu dataset and reported a MOS of 3.76. Naseem et al. (2025) trained Tacotron 1 and Tacotron 2 systems on Urdu and Punjabi data (8–14k utterances, $\leq$10 seconds per utterance), they showed significant improvements over Meta’s MMS baseline. But their data was not released publicly. To mitigate data scarcity, some studies employed transfer learning. Jamal et al. (2022) fine-tuned English and Arabic Tacotron models on Urdu. They found that cross-lingual transfer improves MOS ($\approx$3.3–3.4) compared to scratch training ($\approx$2.9). Hanzala and Kanwal (2024) adapted the multi-speaker Tortoise-TTS model on 16k Urdu samples and achieved a MOS score between 3.6 and 4.2. They also reported that multi-speaker fine-tuning outperforms single-speaker setups. Khan et al. (2024) fine-tuned the multilingual SpeechT5 model on Mozilla Foundation (2020) Urdu dataset which is a publicly available dataset, though data quality and pronunciation issues persisted.

2.3. Dataset Development

Availability of Dataset remains a major challenge in Urdu TTS research. Habib et al. (2014) proposed a greedy selection algorithm to construct an 8-hour phonetically rich dataset covering over 99% of frequent unigrams and bigrams, but this dataset was not released publicly. Adeeba et al. (2016) also created a 10-hour corpus, but it is still unavailable. Jamal et al. (2022) released multiple small datasets (1–11.6 hours) on GitHub, some of them included synthetic Google TTS speech with limited naturalness. Hanzala and Kanwal (2024) collected 16k utterances but did not release their data. Ai4Bharat released 71 hours of Urdu ASR data Javed et al. (2024). The dataset is open-sourced, but it suffer from high variability and un-

Table 1: Data sample, English translation "Whose example is impossible among the slaves of the self."

Phonemized	Romanized	Urdu
ɟɪs ki mɪs'aːl bˌəndəg'aːn n'afs m'ẽ n'aːmʊmkˌɪn hɛ	JS KY MSAL BNDGAN NFS MYN NAMMKN HE	جس کی مثال بندگانِ نفس میں ناممکن ہے

even quality similar to Mozilla Common Voice data. This is likely due to its collection through crowdsourcing, as in IndicSUPERB benchmark (Javed et al., 2023). Naseem et al. (2025) reported substantial 20-hour dataset with phonetic annotation. But their dataset's release status is still unclear. Standardized, high-quality quality noise-free, publicly accessible datasets are still a challenge.

2.4. Romanized vs. Phonemized Script Representations

Another important dimension for good TTS systems is input representation. Most neural Urdu TTS research has relied on Romanized Urdu, which lacks orthographic standardization (Ahmed, 2009). Due to this, crucial phonetic contrasts such as aspiration (/k/ vs. /k^h/), vowel length, retroflexion, and nasalization are lost (Hussain, 2004). Butt et al. (2025) evaluated Roman-Urdu transliteration with transformer models and achieved ≈96% character-level BLEU. Even with the use of advanced models, the systematic errors were not mitigated.

Phonetic transcription provides a precise mapping of Urdu's 44 consonants and complex vowel system. Hussain (2004) established letter-to-sound rules mapping for the Perso-Arabic script to CISAMPA phonemes which enables consistent pronunciation modeling. Traditional systems (Adeeba et al., 2016; Habib, 2014) have tried to use such phonetic transcriptions earlier. Modern neural systems often avoid them due to the complexity of lexicon construction. Jamal et al. (2022) and Saba et al. (2022) relied exclusively on Romanized input while Khan et al. (2024) used raw Urdu text without explicit phonetic modeling.

No previous work has directly compared the importance of Phonemized representation over the Romanized one, even with the theoretical understanding of the advantages of the phonetic inventories. This gap motivates our present study, where we try to empirically evaluate both approaches through objective and subjective metrics on the benchmark dataset.

3. Our Contributions

To overcome these challenges, we make the following contributions in this work:

- **Dataset creation:** We introduce URDUTTS, an 89-hour studio-quality speech dataset at 22.05 kHz. The dataset includes both mono-speaker and multi-speaker recordings (four speakers in total), and provides transcriptions in three formats: native Urdu script, Romanized Urdu script, and Phonemized Urdu script. To our knowledge, this is the largest high-quality dataset with 3 way transcriptions ever created for Urdu TTS.

- **Benchmarking with modern models:** We benchmark this dataset using strong non-autoregressive neural TTS models, namely VITS (Kim et al., 2021) and Glow-TTS (Kim et al., 2020). These models have proven state-of-the-art for many languages, and here we adapt them to systematically evaluate Urdu speech synthesis under the three text representations.

- **Phonemized vs. Romanized Urdu:** We conducted a comprehensive evaluation to test our central hypothesis, that Phonemized Urdu script provides superior pronunciation quality compared to Romanized Urdu. Our evaluation includes both objective metrics (MCD, WER, CER, SNR) and subjective listening tests with 40 native Urdu speakers using MOS, with special emphasis on pronunciation and naturalness.

This work releases a large, publicly available corpus, presents a comprehensive benchmark comparing Phonemized and Romanized transcriptions, and directly addresses the limitations of prior approaches. This study establishes a new resource and evaluation standard for Urdu speech synthesis and highlights the importance of Phonemized transcriptions for other low-resource languages with complex phonetic inventories.

4. Dataset

4.1. Background

Although Urdu is spoken by hundreds of millions of people, high-quality speech resources for research are scarce. Public efforts such as Mozilla Foundation (2020) provide sizable amounts of Urdu audio (tens of hours), but the recording

Table 2: Dataset statistics for each speaker and overall totals.

Speaker	Duration (hrs)	#Samples	Avg Dur. (s)	Avg Words	Avg Chars
Speaker 1	27.38	15,160	6.20	17.00	73.54
Speaker 2	29.95	15,900	6.78	18.36	81.79
Speaker 3	21.56	16,200	4.79	12.74	53.99
Speaker 4	9.70	8,190	4.26	10.99	49.76
Total / Avg	88.59	55,450	5.50	14.77	64.77

quality and annotation consistency vary substantially. Prior academic efforts typically released only small data (on the order of a few hours) or did not publish their data under an open research license as discussed in subsection 2.3, which limits reproducibility and progress in Urdu speech and language technologies. To address this, we present **URDUTTS**, an 89-hour, studio-quality Urdu TTS dataset recorded at 22.05 kHz (16-bit PCM). The dataset contains both mono-speaker and multi-speaker material (four speakers in total) and provides three aligned text representations for every utterance: (1) native Urdu script, (2) Romanized Urdu script, and (3) Phonemized Urdu script as shown in Table 1. Our aim is to release URDUTTS so that it can serve as a benchmark for TTS, ASR and phonetics research.

4.2. Dataset Collection

All source audio was obtained from publicly available recordings (such as LibriVox, internet archive, Loyal Books etc) that met an initial quality filter to ensure speaker consistency, pronunciation quality, no audible background noise, and consistent recording conditions. Recordings were standardized to 22.05 kHz, 16-bit Pulse Code Modulation (PCM) before further processing.

4.2.1. Audio Segmentation

Long continuous recordings were segmented into short clips suitable for model training. Segmentation was automated using the `pydub` library's silence detection with the following heuristic:

- Detect a silent interval if the pause is greater than 400 ms and the amplitude is below -15 dB.
- Do not accept clips shorter than 2 seconds.
- Discard clips longer than 15 seconds at this stage; only clips with final length between 2 and 15 seconds were retained for the corpus.

With this empirical setup, manual clipping was avoided, and further evaluation of the audio were taken into account in further section.

4.2.2. Segment Synchronization

Automatic segmentation of audio clips can cut at sharp boundaries, sometimes splitting lexical items across adjacent clips. To reduce unnatural truncation and preserve word onsets/offsets, we applied a light synchronization heuristic, by empirically trimming 100 ms from the end of clip i and prepended the same 100 ms to the start of clip $i+1$ when both clips originated from the same contiguous recording session. Hence, boundaries are smoothen and abrupts are reduced with preserving the original continuous timing.

4.2.3. Automatic Transcription and Post-editing

Automatic Urdu script transcriptions were produced with a Whisper-based pipeline implemented using the `faster-whisper` (Radford et al., 2023) wrapper to generate time-aligned text. Although Whisper has its own limitations (Sehar et al., 2025), it is currently accepted for Urdu ASR as there are no better well documented open-source ASR available to the community. The transcripts were then converted into Romanized and Phonemized texts. **Romanized Transcriptions** were generated by a deterministic transliteration module (Ahmed, 2009) that maps Urdu graphemes to Latin characters according to a fixed convention. The **Phonemized transcriptions** were generated by the `espeak-ng` (eSpeak NG Project, 2015) toolkit to produce phoneme sequences for each utterance.

4.3. Intended Use and Release

URDUTTS is intended as a public benchmark for TTS, ASR, and phonetics research in Urdu language. The dataset is available on our project GitHub and Hugging Face under an explicit research license (details and the exact license text will be provided on the repository). The release will include:

- 89 hours audio clips (22.05 kHz, 16-bit PCM),
- monospeaker and multispeaker format,
- Urdu transcriptions,

- corresponding Romanized and Phonemized transcriptions,
- codes for dataset construction and model training for reproducibility,
- metadata files and a short usage guide describing recommended train/test splits and citation instructions.

Table 3: MOS rating scale combining pronunciation and naturalness.

Score	Description
5	Fully natural, no mispronunciations
4	Mostly natural, minor errors, understandable
3	Somewhat unnatural, noticeable errors but intelligible
2	Unnatural, frequent errors, difficult to follow
1	Completely unnatural, unintelligible

4.4. Reproducibility

All scripts used for segmentation, synchronization, normalization, transliteration, phonemization and metadata generation (including configuration for `pydub`, the `faster-whisper` pipeline, and `espeak-ng` settings) will be released alongside the dataset to ensure reproducibility and to allow other researchers to adapt the pipeline for different datasets or languages.

4.5. Dataset Quality

To assess the quality of the corpus, both objective and subjective evaluations were carried out. We computed the Voice Activity Detection–based Signal-to-Noise Ratio (VAD-SNR), which yielded an average score of **48.52** dB across the dataset, as mentioned in Table 4. This confirms that the audio recordings have very low noise levels and high acoustic clarity. A listening test was conducted with 40 native Urdu speakers. Participants rated the recordings of the dataset on a 5-point Mean Opinion Score (MOS) scale, with emphasis on naturalness and pronunciation quality. The dataset achieved an average **MOS** of 4.79 ± 0.12 as mentioned in Table 5. In the test, listeners perceived the recordings as highly natural and of excellent quality.

4.6. Dataset Statistics

The URDUTTS dataset contains approximately 89 hours of recorded speech at a sampling rate of 22.05 kHz and 16-bit resolution. The dataset is composed of more than 55K audio clips, and all speakers are male, further statistics are mentioned in Table 2.

5. Experimental Setup

5.1. Models and Training Setup

We evaluated a range of both autoregressive and non-autoregressive TTS models, including Tacotron 2 (Shen et al., 2018), TransformerTTS (Li et al., 2019), Neural HMM (Mehta et al., 2022), VITS (Kim et al., 2021), Glow-TTS (Kim et al., 2020), DelightfulTTS (Liu and et al., 2021), and FastSpeech (Ren et al., 2019). Based on preliminary experiments, we observed that non-autoregressive models, particularly VITS and Glow-TTS consistently produced higher-quality audio while also offering faster training convergence and more efficient inference. Consequently, these models were selected as our primary systems for benchmarking.

Each model was trained twice, once using the Romanized text representation and once using the Phonemized text representation. All experiments were conducted using the `CoquiTTS` framework (Coqui AI, 2020), with models trained entirely from scratch. Training was performed on an NVIDIA V6000 GPU with 48 GB of RAM, using a batch size of 64 for 3,000 epochs. The single-speaker 1 subset of the URDUTTS dataset was used for training in all cases.

Training VITS required approximately 5 days on Romanized text and 6 days on Phonemized text, while Glow-TTS required about 7 days and 8 days, respectively. A fair comparison between Romanized and Phonemized text representations was conducted.

5.2. Evaluation Metrics

To thoroughly evaluate both the performance of the trained models and the quality of the URDUTTS dataset, we adopted a comprehensive evaluation strategy that combines subjective and objective metrics so that both human perception and quantitative acoustic characteristics are captured.

5.2.1. Subjective Evaluation

For the subjective evaluation, a total of 250 audio samples were selected, 50 from each system (ground truth reference, VITS-Phonemized, VITS-Rmanized, GlowTTS-Phonemized and GlowTTS-Rmanized). These were evaluated by 40 native Urdu speakers in a controlled, noise-free environment. Participants were instructed to rate each sample using the Mean Opinion Score (MOS) protocol Table 3, with scores ranging from 1 (bad) to 5 (excellent). Special emphasis was placed on two aspects critical to Urdu: (i) pronunciation accuracy, particularly of nasalized vowels, aspirated

Table 4: Objective evaluation of systems. 500 samples were evaluated for each system. Ph and Rm acronyms are used for Phonemized and Romanized Urdu scripts respectively.

System	**MCD Plain ↓**	**MCD DTW ↓**	**MCD DTW-SL ↓**	**SNR (dB) ↑**	**WER (%) ↓**	**CER (%) ↓**
VITS-Ph	17.00 ± 2.71	7.97 ± 1.66	9.35 ± 2.19	61.86	12.3	5.3
VITS-Rm	17.53 ± 2.96	8.53 ± 1.74	9.98 ± 2.26	35.39	15.1	6.5
Glow-Ph	17.71 ± 2.01	6.74 ± 0.98	7.42 ± 1.26	10.40	12.5	5.8
Glow-Rm	18.83 ± 2.12	7.07 ± 1.07	7.86 ± 1.44	10.27	25.0	11.8

plosives, and retroflex sounds, and (ii) naturalness of speech. MOS values were reported with 95% confidence intervals. There are limitations of MOS (Le Maguer et al., 2024), but it is a well accepted evaluation criteria in speech synthesis tasks.

5.2.2. Objective Evaluation

For the objective evaluation, 500 audio samples were analyzed for each system and the ground truth. The following metrics were employed:

- **Mel-Cepstral Distortion (MCD):** Measures the distance between the mel-cepstral coefficients of generated and reference audio (Kubichek, 1993; Tokuda et al., 2000; Kim et al., 2020, 2021). We report three variants:

  1. **MCD-Plain** measures frame-by-frame spectral differences. It reflects the overall spectral fidelity, but is sensitive to slight temporal misalignments.
  2. **MCD-DTW** applies Dynamic Time Warping to account for small misalignments in speech timing, which often occur in synthesized outputs. This ensures that evaluation focuses on spectral similarity rather than timing mismatch.
  3. **MCD-DTW_SL** gives a slope limitation in DTW to prevent over-flexible alignment, giving a more realistic measure of pronunciation accuracy and prosody.

  Lower MCD values indicate better spectral similarity (Kubichek, 1993).

- **Signal-to-Noise Ratio (SNR):** We use Voice Activity Detection (VAD)-based SNR to estimate the ratio of speech energy to background noise. Higher SNR values indicate cleaner and higher-quality audio.

- **Word Error Rate (WER):** Measures the percentage of word-level mismatches between the ASR-transcribed generated output and the reference transcript. Computed as:

$$\text{WER} = \frac{S + D + I}{N} \times 100$$

  where S = substitutions, D = deletions, I = insertions, and N = total words. Lower WER indicates higher intelligibility.

- **Character Error Rate (CER):** Similar to WER but computed at the character level, making it sensitive to pronunciation errors in morphologically rich languages such as Urdu. Defined as:

$$\text{CER} = \frac{S + D + I}{N} \times 100$$

  where the operations are measured on characters instead of words.

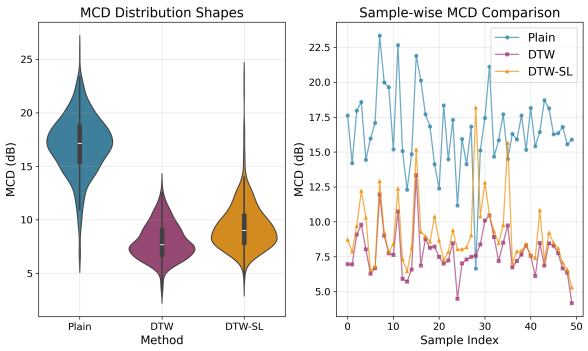
Jointly analyzing subjective MOS scores and objective metrics (MCD, SNR, WER, CER), our evaluation provides one of the most comprehensive and fine-grained assessments of Urdu TTS to date. We thus evaluate spectral fidelity, noise, intelligibility and pronunciation accuracy. It ensures that both human-perceived naturalness and machine-measured fidelity are rigorously benchmarked.

6. Results

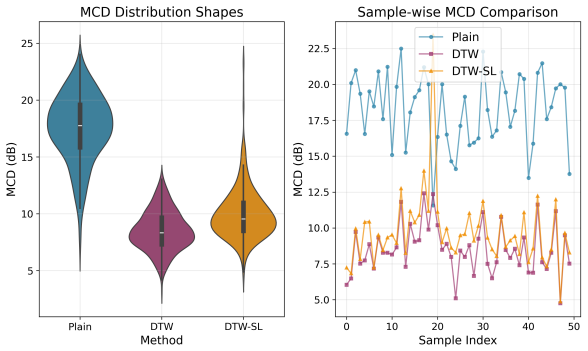
The evaluation results demonstrate a clear and consistent advantage of using Phonemized Urdu input over Romanized text for both VITS and GlowTTS models.

Subjective evaluation: As shown in Table 5, VITS-Phonemized achieves the highest MOS, closest to the natural speech reference which indicates that listeners perceive it as highly natural and accurate in pronunciation. GlowTTS-Phonemized also outperforms its Romanized counterpart, though both GlowTTS models score lower than VITS.

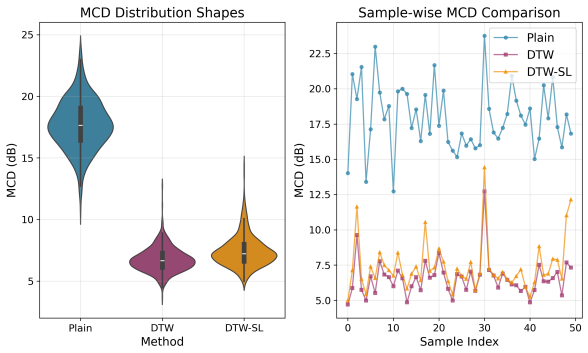
Objective evaluation: 500 samples were evaluated per model. Table 4 shows that Phonemized models consistently achieve lower MCD values which reflects better spectral fidelity and more accurate phonetic realization. Figure 1 and Figure 2 shows the statistics of MCD metrics and variability over the test audio samples. Similarly, VAD-SNR values in Table 4 indicate that Phonemized VITS outputs are cleaner, with less background noise or artifacts.



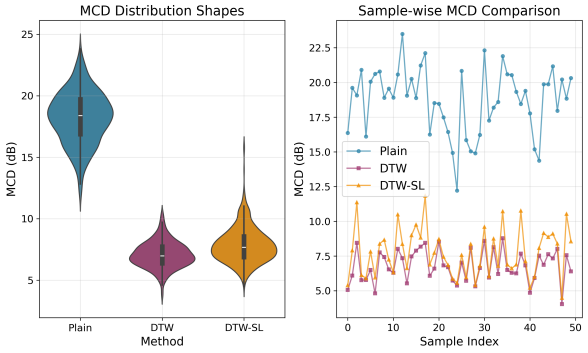
(a) VITS-Phonemized



(b) VITS-Rmanized

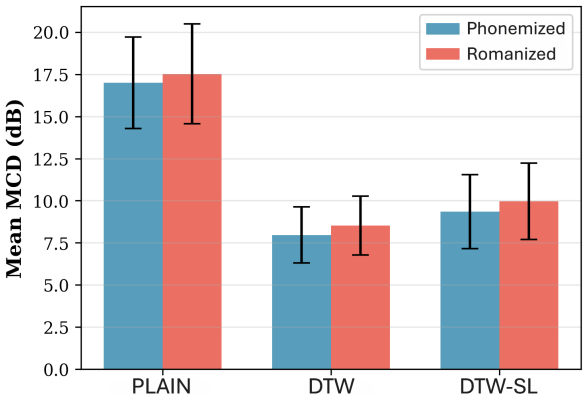


(c) GlowTTS-Phonemized

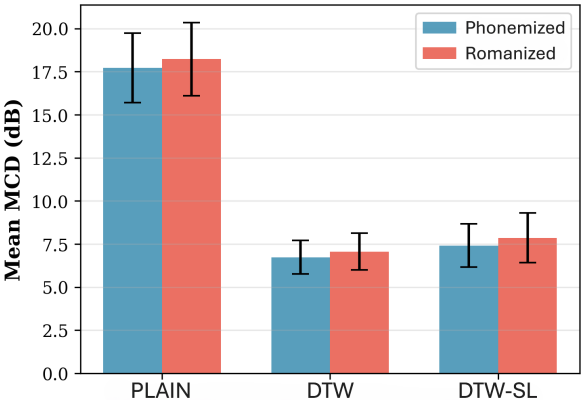


(d) GlowTTS-Rmanized

Figure 1: MCD statistics for VITS (a) Phonemized, (b) Romanized and GlowTTS (c) Phonemized, (d) Romanized. Left: violin plots show the distribution of Plain, DTW, and DTW-SL. Right: Sample-wise trends highlight variability, with Plain exhibiting larger fluctuations than others.



(a) VITS



(b) GlowTTS

Figure 2: MCD statistics for (a) VITS Phonemized vs. Romanized, (b) GlowTTS Phonemized vs. Romanized.

Table 5: Mean Opinion Score (MOS). 50 samples were evaluated for each system. Scores include 95% confidence intervals.

System	MOS
Reference	4.79 ± 0.12
VITS-Phonemized	4.38 ± 0.24
VITS-Rmanized	3.86 ± 0.36
GlowTTS-Phonemized	3.34 ± 0.28
GlowTTS-Rmanized	2.64 ± 0.31

Word and character error rates were taken after transcribing audio generated by models via `faster-whisper` model. In Table 4, WER and CER are reported which provides an explicit measure of intelligibility. Phonemized systems achieve substantially lower WER and CER than Romanized ones. It gives the understanding that phonetic representation reduces pronunciation errors and enhances comprehension.

Both subjective and objective results confirm that Phonemized Urdu consistently outperforms Romanized counterpart across all metrics. The Phonemized systems produce speech that is

Table 6: Top-5 hardest and easiest words by recognition accuracy for each model.

Hardest Words			Easiest Words		
Model	**Word**	**Acc.**	**Model**	**Word**	**Acc.**
VITS-Ph	جدہ	14.3%	VITS-Ph	عزت	100%
	عنہ	20.0%		زبان	100%
	استغراق	20.0%		تجھ	100%
	پناہ	40.0%		صحت	100%
	پروانہ	40.0%		کس	100%
VITS-Rm	صحت	0.0%	VITS-Rm	عزت	100%
	بڑی	16.7%		مقصد	100%
	رضی	20.0%		زبان	100%
	عنہ	20.0%		تجھ	100%
	استغراق	20.0%		ذات	100%
GlowTTS-Ph	استغراق	0.0%	GlowTTS-Ph	زبان	100%
	جدہ	28.6%		تیرا	100%
	پروانہ	40.0%		تجھ	100%
	صحت	42.9%		لوگ	100%
	امیر	57.1%		مکان	100%
GlowTTS-Rm	صحت	0.0%	GlowTTS-Rm	عزت	100%
	بڑی	16.7%		مقصد	100%
	رضی	20.0%		زبان	100%
	عنہ	20.0%		تجھ	100%
	استغراق	20.0%		ذات	100%

Table 7: Top substitution confusions (reference → hypothesis) per model.

Model	Reference → Hypothesis	Count	Linguistic Note
VITS-Ph	ہو → ہوں	4	Nasalisation loss
	اس → اسے	3	Case-suffix deletion
	کہا → کہاں	3	Nasal coda drop
	خدا → خداونت	3	Word truncation
VITS-Rm	سے → اسے	11	Initial alef deletion
	بری → بڑی	5	Retroflex confusion
	اسے → اس	5	Case-suffix insertion
	برا → بڑا	3	Retroflex confusion
GlowTTS-Ph	اب → اور	17	Conjunction instability
	جدا → جدہ	5	Final ہ deletion
	کہا → کہاں	3	Nasal coda drop
	میر → امیر	3	Initial alef deletion
GlowTTS-Rm	سے → اسے	11	Initial alef deletion
	بری → بڑی	5	Retroflex confusion
	اسے → اس	5	Case-suffix insertion
	فلاں → فلاں	3	Nun-ghunna vs. noon

more natural, clearer, and closer to human pronunciation. Hence, it validates the effectiveness of phonetic transcription in improving Urdu TTS quality.

7. Linguistic Analysis

To better understand model behaviour beyond aggregate metrics, we conducted a word-level linguistic analysis across all four TTS systems. For each model, we tracked per-word recognition ac-

Table 8: Mean recognition accuracy (%) by word length (Unicode characters).

Model	2	3	4	5	6
VITS-Ph	94.9	90.3	93.5	93.9	91.7
VITS-Rm	89.8	84.8	92.2	94.3	91.5
GlowTTS-Ph	96.4	92.6	96.9	98.1	86.2
GlowTTS-Rm	89.8	84.8	92.2	94.3	91.5

curacy (the fraction of utterances in which a reference word was reproduced correctly by the ASR transcription), identified the most common substitution confusions, and examined how word length relates to intelligibility. The analysis filters words appearing fewer than five times to ensure statistical reliability.

7.1. Hardest and Easiest Words

Table 6 reports the five words with the lowest and highest recognition accuracy for each model. The word استغراق (*istighrāq*, "absorption/contemplation") is consistently difficult across all models, reflecting both its low corpus frequency and complex phonetic structure. جدہ (*jaddah*) and عنہ (*'anhu*) are similarly problematic. By contrast, high-frequency, phonetically simple words such as زبان (*zubān*, "language") and تجھ (*tujh*, "you") achieve 100% accuracy across all systems, suggesting that these forms are robustly synthesised regardless of the acoustic model or input script.

7.2. Common Substitution Confusions

Table 7 lists the most frequent word-level substitution pairs per model. Several patterns recur across systems. The confusion اسے → سے (pronoun "it/him" reduced to postposition "from") is the dominant error in both romanised models (11 occurrences each), pointing to a systematic difficulty in distinguishing word-initial ا in short function words. Similarly, جدہ → جدا appears in both phonemised models, suggesting that the word-final ہ is consistently under-synthesised. The GlowTTS-Ph model exhibits a notable اور → اب confusion (17 occurrences), which may indicate instability in synthesising the common conjunction اور under that architecture. The pair بڑی → بری ("big" vs. "bad") appearing in both romanised models reflects a well-known Urdu minimal pair distinguished only by the retroflex ڑ, which is challenging for models trained on romanised input.

7.3. Word Length and Recognition Accuracy

Table 8 shows mean recognition accuracy grouped by word length (in Unicode characters) for each model. Contrary to the general expectation that longer words are harder to synthesise, the results show a non-monotonic pattern: 3-character words consistently exhibit the lowest accuracy, while 5-character words often perform best. This likely reflects the high proportion of short function words (pronouns, postpositions, conjunctions) at lengths 2–3 that are phonetically similar and easily confused, rather than a simple length–difficulty relationship. All models show similar trends, suggesting this is a corpus-level property rather than an architectural one.

Overall, the linguistic analysis reveals that model errors cluster around (i) phonetically ambiguous short function words, (ii) words containing retroflex consonants (ڑ, ڈ, ٹ) which are absent from Arabic and therefore potentially underrepresented in model priors, and (iii) rare lexical items with complex morphology. Phonemized models show a qualitative advantage in handling the retroflex distinction, while Romanized models suffer more from function word confusions involving word-boundary effects, and overall Phonemized performs significantly well than Romanized, backing our hypothesis (Phonemized are better for speech synthesis than Romanized).

8. Conclusion

In this work, we addressed the challenge of developing a high-quality TTS system for Urdu, which is a language with rich phonetic diversity but scarce computational resources. We introduced **URDUTTS**, an 89-hour studio-quality, multi-format dataset(native Urdu script, Romanized Urdu, and Phonemized Urdu), with both mono-speaker and multi-speaker configuration, which to our knowledge is the largest open dataset of its kind for Urdu speech synthesis. Through extensive experiments with modern non-autoregressive TTS models (VITS and GlowTTS), we demonstrated that Phonemized Urdu input consistently outperforms Romanized input, achieving higher naturalness, intelligibility, and lower distortion across both subjective (MOS) and objective (MCD, SNR, WER, CER) evaluations and with linguistic analysis. This confirms our central hypothesis that phonetic representations are crucial for handling Urdu's complex sound system and for producing natural-sounding speech.

Beyond benchmarking, our results highlight broader implications: (1) Phonetic transcription should be prioritized over Romanization for low-resource languages with rich phonologies; (2) A large-scale, high-quality dataset like URDUTTS provides a solid foundation for downstream tasks such as Automatic Speech Recognition (ASR), and linguistic research.

9. Data and Code Availability

We release URDUTTS with a permissive license to encourage open research and reproducibility. We hope that this resource and our findings will serve as a benchmark for future Urdu TTS work, and also as a model for developing speech resources in other underrepresented languages. The data and code are available at github.com/KAABSHAHID/URDUTTS.

10. Ethics Statement

The URDUTTS dataset was curated with careful consideration of ethical principles. We ensured that no personally identifiable information (PII) is contained in the dataset, and all contributions were anonymized before release. The dataset is intended strictly for research and educational purposes in the fields of computational linguistics and speech technology.

References

Farah Adeeba et al. 2016. Comparison of urdu text to speech synthesis using unit selection and hmm based techniques. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Tafseer Ahmed. 2009. Roman to urdu transliteration using wordlist. In *Proceedings of the Conference on Language and Technology*, volume 305. ELRA.

M.S. Amin, X. Zhang, L. Anselma, and et al. 2025. Semantic processing for Urdu: corpus creation, parsing, and generation. *Language Resources & Evaluation*, 59:2469–2500.

S. Arif, A. J. Khan, M. Abbas, A. A. Raza, and A. Athar. 2025. WER we stand: Benchmarking Urdu ASR models. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 5952–5961, Abu Dhabi, UAE. Association for Computational Linguistics.

Elena Bashir. 2011. Urdu and linguistics: A fraught but evolving relationship.

Umer Butt, Stalin Veranasi, and Günter Neumann. 2025. Low-resource transliteration for roman-urdu and urdu using transformer-based models. *arXiv preprint arXiv:2503.21530*.

Coqui AI. 2020. Coqui tts: A deep learning toolkit for text-to-speech. https://github.com/coqui-ai/TTS. Accessed: 2025-10-09.

eSpeak NG Project. 2015. espeak ng: An open source speech synthesizer. https://github.com/espeak-ng/espeak-ng. Accessed: 2025-10-09.

Tania Habib. 2014. Hidden markov model (hmm) based speech synthesis for urdu language. In *Proceedings of the Conference on Language and Technology*, Islamabad, Pakistan. Center for Language Engineering (CLE).

Wajiha Habib et al. 2014. Design of speech corpus for open domain urdu text-to-speech system using greedy algorithm. In *Conference on Language and Technology (CLT14)*.

Ahmed Hanzala and Ayesha Kanwal. 2024. Generative urdu speech synthesis. In *2024 4th International Conference on Computer Systems (ICCS)*. IEEE.

Sarmad Hussain. 2004. Letter-to-sound conversion for urdu text-to-speech system. In *Workshop on Computational Approaches to Arabic Script*.

Sahar Jamal, Sadaf Abdul-Rauf, and Quratulain Majid. 2022. Exploring transfer learning for urdu speech synthesis. In *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia, 13th Language Resources and Evaluation Conference (LREC 2022)*. ELRA.

Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M. Khapra. 2023. Indicsuperb: a speech processing universal performance benchmark for indian languages. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

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, Vaijayanthi C, Krishnan Karunganni, Pratyush Kumar, and Mitesh Khapra. 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.

Salman Ahmad Khan, Musadaq Mansoor, and Abdullah Habib. 2024. Overcoming linguistic barriers: Developing advanced urdu text-to-speech systems. In *2024 19th International Conference on Emerging Technologies (ICET)*. IEEE.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Robert Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128. IEEE.

Sébastien Le Maguer, Simon King, and Naomi Harte. 2024. The limits of the mean opinion score for speech synthesis evaluation. *Computer Speech & Language*, 84:101577.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Ming Liu, and Ming Zhou. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6706–6713.

Y. Liu and et al. 2021. Delightfultts: The microsoft speech synthesis system for blizzard challenge. In *Proceedings of the Blizzard Challenge Workshop*.

Shivam Mehta, Jonas Külske, Matthias Sperber, Matthias Paulik, and Björn Schuller. 2022. Neural hmms are all you need (for high-quality attention-free tts). In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7897–7901. IEEE.

Mozilla Foundation. 2020. Mozilla common voice: Urdu dataset. https://commonvoice.mozilla.org/en/datasets. Accessed: 2025-09-29.

Fatima Naseem et al. 2025. Developing high-quality tts for punjabi and urdu: Benchmarking against mms models. In *Proceedings of Interspeech 2025*. ISCA.

Sultan Mahmood Niazi and Perveen Akhter Farhat. 2023. Contrastive analysis of the vowel sound system of urdu and english languages. *Pakistan Languages and Humanities Review*, 7(4):228–240.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems*, volume 32, pages 3171–3180.

Riffat Saba, Hikmat Ullah Khan, and Muhammad Bilal. 2022. Urdu text-to-speech conversion using deep learning. In *2022 International Conference on IT and Industrial Technologies (ICIT)*. IEEE.

Najm Ul Sehar et al. 2025. Benchmarking whisper for low-resource speech recognition: An n-shot evaluation on pashto, punjabi, and urdu. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*. Association for Computational Linguistics.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, and Rif A. Saurous. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura. 2013. Speech synthesis based on hidden markov models. In *Proceedings of the IEEE*, volume 101, pages 1234–1252.

Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, and Takao Kobayashi. 2000. Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 88(3):463–497.

M. Usama, S. Mehboob, and R. Waheed. 2024. Linguistic variations in urdu lexicon in everyday communication: A case study of the last decade of the modern century. *Advances in Humanities Research*, 4:25–36.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, and Rob A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Wikipedia contributors. 2025. List of languages by total number of speakers. https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers. [Online; accessed 29-September-2025].