Deep learning-based Voice Cloning for Local Languages with a Focus on Urdu

Mujtaba Mateen
Department of Computer Information & Systems Engineering,
NED University of Engineering & Technology
Karachi, Pakistan
mateen4301991@cloud.neduet.edu.pk

Hunzala Mushtaq
Department of Computer Information & Systems Engineering,
NED University of Engineering & Technology
Karachi, Pakistan
mushtaq4304742@cloud.neduet.edu.pk

M. Usama Shahid
Department of Computer Information & Systems Engineering,
NED University of Engineering & Technology
Karachi, Pakistan
shahid4304218@cloud.neduet.edu.pk

Muneeb Ullah Khan
Department of Computer Information & Systems Engineering,
NED University of Engineering & Technology
Karachi, Pakistan
khan4301814@cloud.neduet.edu.pk.

***Abstract*— Recent advancements in text-to-speech (TTS) and voice cloning have enhanced human-computer interaction, yet the Urdu language, spoken by over 100 million people, remains largely overlooked. Existing solutions often fail to capture the language's nuances. Collaborating with Anjuman Taraqqi-e-Urdu Pakistan (ATUP), This paper aims to address this gap by developing a high-quality voice synthesis model for Urdu, leveraging deep learning techniques and creating a comprehensive dataset, improving digital accessibility and inclusivity for Urdu speakers. The methodology includes data collection, preprocessing, model training, and inferencing, with a focus on the eXtended Text-to-Speech (XTTS) model. Evaluation using the Mean Opinion Score (MOS) showed XTTS achieved a score of 3.14, indicating fair to good naturalness and intelligibility.**

Keywords—Text-to-speech, Speech synthesis, Urdu language, Deep Learning, Voice cloning, XTTS.

I. INTRODUCTION

Recent advancements in communication technologies have significantly expanded global interactions, with modern speech synthesis technologies like Text-to-Speech (TTS) and voice cloning [1] emerging as valuable tools for language learning and cross-lingual communication. However, despite the global linguistic diversity, there is limited technological development for the Urdu language, which has over 100 million speakers. Existing TTS systems often fail to capture the nuances and local dialects of Urdu, resulting in synthesized speech that sounds overly generalized and lacks natural cadence and pronunciation. This deficiency hampers accessibility and relevance for Urdu speakers in the digital age. Furthermore, these is no Urdu speech corpus readily available [2].

Voice cloning [3] plays a crucial role in text-to-speech (TTS) systems, which convert written text into spoken words. Since 2016, deep learning and generative models for synthesizing more natural-sounding speech have emerged, surpassing traditional concatenative methods. Research has since concentrated on enhancing these models to make them sound more natural and to enable end-to-end training. Initially, inference on Graphical Processing Units (GPUs) [4] was significantly slower than real-time on mobile CPUs, but improvements have led to near-human naturalness in speech quality.

The primary aim is to develop a TTS system that can take an Urdu sentence and synthesize it into speech that replicates the speaker's voice. This study proposes a TTS system for the Urdu language utilizing advanced deep learning and generative modelling techniques. The eXtended Text-to-Speech (XTTS) [5] model is employed, trained on a custom dataset comprising over 1,200 high-quality audio clips and approximately 90,000 words extracted from 80 diverse essays. The Mean Opinion Score (MOS), recommended by the International Telecommunication Union (ITU) for assessing TTS speech quality, is used for evaluation. The remainder of the paper is structured as follows: Section II presents an extensive review of existing literature. Section III outlines the proposed research methodology. Section IV details the results and their discussion, while Section V provides concluding remarks.

II. LITERATURE REVIEW

The search for an appropriate dataset for Urdu text-to-speech encountered several challenges. Primary among these was the limited availability of open-source Urdu audio datasets. Additionally, the absence of comprehensive speaker transcripts posed a significant obstacle to creating standardized TTS systems [2]. While some proprietary datasets were identified, they were primarily geared towards emotion recognition rather than TTS applications [6]. Table 1 summarizes the unsuitability of major audio datasets for Urdu TTS development.

TABLE 1: UNSUITABILITY OF MAJOR AUDIO DATASETS FOR URDU TTS DEVELOPMENT

Dataset	Reason for not using it
LJ Speech [7]	It is an English-language dataset consisting of 13,100 short audio clips.
UrbanSound8K [8]	This dataset is designed for audio scene classification tasks rather than speech synthesis.
VCTK [9]	The dataset contains only English speech.
GigaSpeech [10]	The dataset's primary language is English, with a focus on automated speech recognition (ASR) rather than TTS

In the domain of TTS technology, significant strides have been made to enhance speech synthesis quality, making it more natural and accessible across various languages.

Currently, the researchers have produced natural human-like speech by utilizing a variety of TTS approaches. It was found that TTS techniques proposed using Deep Neural Network (DNN) performed better than concatenative and statistical parametric approaches in generating quality speech [11]. Tacotron [12], Deep Voice [13], and ClariNet [14] are some of the most popular DNN-based TTS techniques.

Pitrelli et al. [15] introduced an expressive TTS system aimed at mimicking human speech's natural quality by varying pitch and frequency according to the context, a departure from the monotonic output of early systems.

In addressing the linguistic and phonetic diversity, Mahanta et al. [16] tackled the inadequacy of American English pronunciation in Indian English TTS systems. Through dictionary modifications and the employment of unit selection synthesis and statistical parametric speech synthesis frameworks, they achieved a significant improvement in pronunciation accuracy, with a notable reduction in the error rate by 7.69%.

Mullah et al. [17] furthered this linguistic adaptation by employing HMM in the development of an Indian accent speech synthesis system. The exploration of TTS systems extends beyond English, addressing the scarcity of research in languages like Arabic and Urdu. Bilecen and Arioz [18] expanded the linguistic reach by translating an English TTS system into French.

Neural network-based end-to-end TTS systems represent a significant advancement in the field, eliminating the need for human feature extraction and preprocessing steps required by traditional methods. The advent of WaveNet [4] introduced a neural network-based approach that directly generates waveforms from linguistic features.

Subsequent developments, such as DeepVoice [13] Tacotron [12], and their iterations, have progressively enhanced speech quality and efficiency, incorporating mechanisms like attention and post-processing neural vocoders to improve speech synthesis further.

III. RESEARCH METHODOLOGY

A. *Data Acquisition*

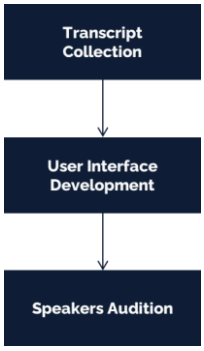

Fig. 1. Data Collection Process

The lack of accessible datasets for Urdu text-to-speech systems underscores a significant challenge, prompting us to create a new dataset. This effort was driven by the compilation of 80 essays across Religious, Pakistan's history, Science, and Abstract topics, sourced from diverse websites like HamariWeb [19] and Rekhta [20]. We also developed a streamlined user interface that facilitated easy audio recording, automated data annotation, and rigorous human verification at each stage as shown in Fig. 2. Auditions conducted at ATUP enabled the selection of speakers with proficient Urdu skills, crucial for our voice cloning project.

TABLE 2: CORPUS STATISTICS

Measures	Our Dataset	
Speakers	Male Speaker (ATUP)	Female Speaker (ATUP)
Total Audio Clips Duration	4.71 hours	5.45 hours
Total Words	43,482	46,019
Mean Clip Duration	67.85 seconds	22.13 seconds
Total Characters	174,755	183,706
Minimum Clip Duration	9.84 seconds	3.78 seconds

Fig. 2. User interface for data collection

B. *Data Preprocessing*

Fig. 3. Data Preprocessing Process

Data preprocessing plays a crucial role in enhancing the quality and naturalness of synthesized speech in TTS systems by cleaning and standardizing audio data. Our approach involved configuring the audio data with a sampling frequency of 44.1 kHz, a 16-bit bit rate, and a mono audio channel, aligning with benchmarks like the LJ Speech dataset. We conducted text normalization and volume normalization to ensure consistency across recordings. Transliteration of Urdu transcripts into Hindi Devanagari script facilitated processing for our voice cloning models, benefiting from shared language features. This transliteration is done using Llama 3 model [21] and prompt engineering.

Noise reduction employed the Spectral Gating technique [22], achieving a STOI score of 92% by effectively reducing various types of noise [23]. Fig. 4 and

Fig. 5 highlights noise reduction and spectrogram analysis, and their impact on signal clarity and quality.

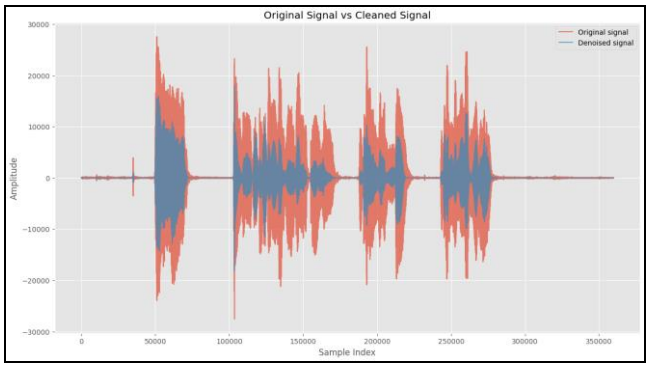

Fig. 4. Audio Signal Analysis

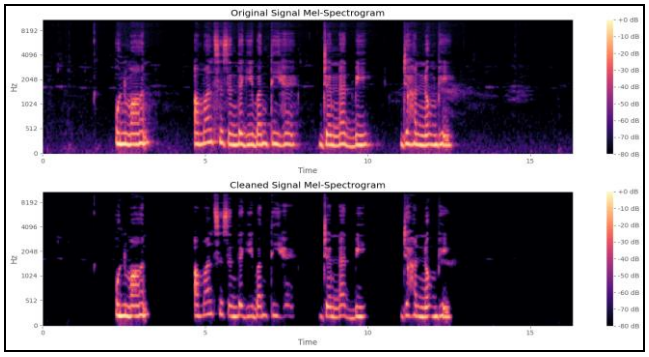

Fig. 5. Mel-Spectrogram Analysis

C. *Model Development*

The TTS systems transform written text into spoken words through several key components, as shown in Fig. 6. The text is analyzed by the feature extractor to determine its phonetic and prosodic properties. The acoustic model then generates audio features based on this linguistic representation, leveraging deep learning techniques for accurate modeling of human speech complexity. Finally, the vocoder converts the acoustic model's output into audible speech, synthesizing natural-sounding speech waves [3]. Vocoder technology has evolved significantly, with newer models using GAN and VAE to capture nuances in human voice.

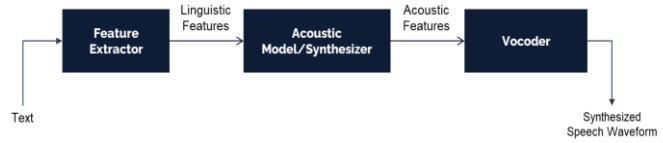

Fig. 6. Key Components of TTS

A transfer learning approach is adopted to leverage state-of-the-art TTS and voice synthesis models developed by industry leaders, guided by selection criteria. XTTS [5] is selected from an initial pool of approximately 50 different TTS models based on these criteria.

The XTTS architecture, detailed in Fig. 7., builds upon Tortoise [24], incorporating modifications for multilingual training and improved performance. It consists of three main components:

- **VQ-VAE:** This Vector Quantized-Variational AutoEncoder, with 13M parameters, encodes mel-spectrogram frames using an 8192-code codebook at a 21.53 Hz frame rate.

- **Encoder:** The Generative Pre-Trained Transformer (GPT-2) based encoder, with 443M parameters, processes text tokens from a custom Byte-Pair Encoding (BPE) tokenizer with 6681 tokens and predicts VQ-VAE audio codes. It is also conditioned by a Conditioning Encoder that converts mel-spectrograms into 32 1024-dimensional embeddings. This Conditioning Encoder, composed of six 16-head Scaled Dot-Product Attention layers and a Perceiver Resampler, offers improved performance.

- **Decoder:** Based on the HiFi-GAN vocoder [25] with 26M parameters, the decoder reconstructs audio from the GPT-2 encoder's latent vectors instead of VQ-VAE codes to avoid pronunciation issues and artifacts. It includes speaker embeddings from the H/ASP model and incorporates Speaker Consistency Loss (SCL) to enhance speaker similarity.

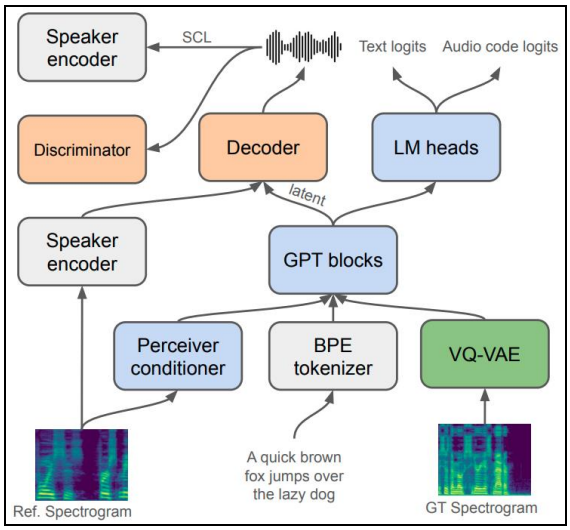

Fig. 7. XTTS Architecture Overview

Training of XTTS is simplified using a Gradio demo, facilitating preprocessing, training, on Nvidia GeForce 1660 SUPER GPU. Coqui TTS library is used for easy training and fine-tuning.

Model inferencing for XTTS involves the loading of pre-trained model configurations and checkpoints, utilizing a FastAPI-based web service for text-to-speech synthesis. The process initializes the model from configuration files and checkpoints, setting up an inference web service with a defined POST endpoint. Urdu text is transliterated to Hindi script, and audio output is generated based on specified parameters, ensuring efficient synthesis using a P100 GPU.

IV. Results

The performance evaluation of the XTTS model utilized the Mean Opinion Score (MOS) metric to gauge the naturalness and intelligibility of synthesized speech. A total of 18 participants assessed audio outputs on a scale from 1 to 5 (excellent, good, fair, poor, and bad), as illustrated in Fig. 8.

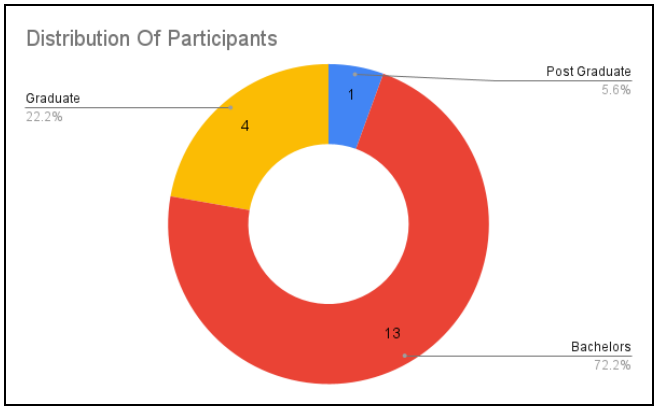

Fig. 8. Distribution of Participants

Survey questions (SQ), detailed in Table 3, encompassed aspects such as overall quality, speech clarity, naturalness, and suitability for learning purposes among non-native Urdu speakers.

TABLE 3: SURVEY QUESTIONS

SQ1	How would you rate the overall quality of the synthesized AI speech?
SQ2	How easily can you understand the speech?
SQ3	How natural does the speech sound to you?
SQ4	How clear is the speech in terms of pronunciation?
SQ5	How is the rhythm and pace of the speech?
SQ6	Are the generated audios being grammatically and sequentially, correct?
SQ7	Can we use it for learning purposes? (Learning Urdu for those people who are not a native speaker of Urdu)
SQ8	Did you hear any background noise in the audios?

The calculated overall MOS for XTTS was 3.14, reflecting a fair to good perception of speech naturalness and intelligibility. The average MOS for each SQ is given in Fig. 9.

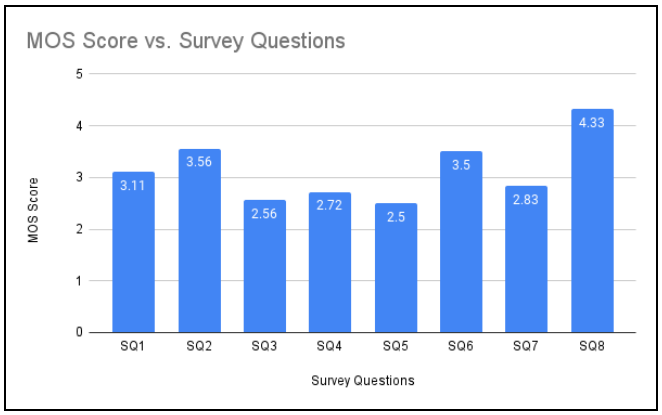

Fig. 9. Average MOS distribution for each SQ

SQ1: This question provides an overall assessment of the synthesized speech's quality from the listener's perspective. It encapsulates subjective impressions of factors like clarity, naturalness, and fidelity to human speech.

SQ2: Understanding speech is fundamental to effective communication. This question evaluates the intelligibility of the synthesized speech, indicating how well the TTS system conveys the intended message without ambiguity or misunderstanding.

SQ3: Naturalness is critical for user acceptance and engagement. This question assesses how closely the synthesized speech mimics natural human speech in terms of intonation, expression, and cadence. Natural-sounding speech enhances user experience and makes interactions with TTS systems more pleasant and effective.

SQ4: Accurate pronunciation is essential for conveying meaning accurately and avoiding misinterpretations. This question evaluates the TTS system's ability to pronounce words and phrases correctly, ensuring that the synthesized speech is clear and understandable.

SQ5: Rhythm and pace influence the flow and comprehensibility of speech. This question examines whether the TTS system delivers speech at an appropriate speed and rhythm, which impacts listening comprehension and user engagement.

SQ6: Grammatical correctness ensures that the synthesized speech follows syntactic rules and is coherent. Sequential correctness ensures that the order of words and phrases in the synthesized speech aligns logically, maintaining clarity and meaning.

SQ7: This question explores the educational utility of the TTS system. It assesses whether non-native speakers can effectively use the system to learn Urdu pronunciation and language structure, indicating its potential for educational applications.

SQ8: Background noise can detract from the clarity and quality of synthesized speech. This question identifies any potential distractions or interference in the audio.

Comparison with existing Urdu TTS models, detailed in Table 4.

TABLE 4: COMPARISON WITH EXISTING URDU TTS

Models	**MOS**
Tacotron (Sahar Jamal)	2.75
Tacotron 2 + WaveGlow (Saba, 2022)	3.76
XTTS (Our Model)	3.14

V. CONCLUSION

This study has focused on advancing Urdu language technology by creating a Text-to-Speech (TTS) system using the XTTS model. We started by carefully gathering a wide range of Urdu speech samples to ensure we had diverse and high-quality data. Our preprocessing work included identifying key features, converting text into Hindi Devanagari script for better learning, and using advanced techniques to reduce background noise in audio, achieving an impressive 92% reduction.

The importance of this project lies in its potential to improve how Urdu speakers interact with computers. Historically, this group has not benefited much from technological advancements. Our Mean Opinion Score (MOS) of 3.14 shows that our system produces speech that is fairly natural and understandable. This success not only helps preserve and promote Urdu but also makes digital interactions more accessible and inclusive.

Looking ahead, we have several ideas to make our Urdu text-to-speech system even better. First, integrating diacritics (Ahrab) into our dataset could improve pronunciation accuracy. Second, refining the model to directly process Urdu text without transliterating it into Hindi would make the system more user-friendly for Urdu speakers. Lastly, partnering with studios to get cleaner audio data could further enhance our dataset, reducing noise during training. These steps aim to strengthen our text-to-speech technology for Urdu, pointing to promising directions for future research and development efforts.

VI. ACKNOWLEDGMENT

We would like to express our gratitude to Almighty Allah for granting us the strength to navigate through challenges and complete half of our final year project on time. We also extend our heartfelt thanks to our Supervisor, Ms. Fauzia Yasir, whose unwavering support and guidance have been instrumental in our progress. Additionally, we are grateful for the guidance of our Co-Supervisor, Dr. Majida Kazmi. Their mentorship has been invaluable, and we are fortunate to have had the opportunity to learn from them. Finally, we thank our families for their constant support and prayers.

REFERENCES

[1] P. M. Patel, "Urdu Zaban And New Technology Issues And Prospectives," New Delhi, 2017.

[2] Tayyaba Fatima, Raees Ul Islam, and Muhammad Waqas Anwar, "Morphological and Orthographic Challenges in Urdu Language Processing: A Review," Department of Computer Science, COMSATS Institute of Information Technology, Lahore, Pakistan.

[3] P. Taylor, "Text-to-speech synthesis," Cambridge University Press, 2009.

[4] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. CoRR, abs/1609.03499, 2016. URL http://arxiv.org/abs/1609.03499.

[5] E. Casanova and K. D. Knight, "XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model," 2024.

[6] I. P. Mushtaq Ahmed, "Urdu Zaban and New Technology: Issues and Prospects," 2017.

[7] K. Ito and L. Johnson, "LJ Speech Dataset," 2017. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[8] C. S. Jacoby, "UrbanSound8K," 2014. [Online]. Available: https://zenodo.org/records/1203745

[9] V. et al., "VCTK," 2016. [Online]. Available: https://paperswithcode.com/dataset/vctk

[10] G. Chen, S. C., "GigaSpeech," 2021. [Online]. Available: https://arxiv.org/abs/2106.06909

[11] O. Watts, G. E., "From HMMs to DNNs: Where Do the Improvements Come From?" in IEEE International Conference on Acoustics, Speech and Signal Processing, 2016, pp. 5505–5509.

[12] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," 2017.

[13] W. Ping, K. Peng, A. Gibiansky, S. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," 2017.

[14] W. Ping, K. Peng, A. Gibiansky, S. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Clarinet: Parallel wave generation in end-to-end text-to-speech," 2018.

[15] J. F. Pitrelli and R. B. Lot, "The IBM expressive text-to-speech synthesis system for American English," IEEE Transactions on Audio, Speech, and Language Processing, pp. 1099-1108, 2006.

[16] D. Mahanta and B. Sarmah, "Text to speech synthesis system in Indian English," in IEEE Region 10 Conference, pp. 2614-2618, 2016.

[17] H. U. Mullah and F. Pathan, "Development of an HMM-based speech synthesis system for Indian English language," in International Symposium on Advanced Computing and Communication, pp. 291-292, 2015.

[18] S. Bilecen and U. Akgul, "Interpretation of English text to speech application to French language," in 24th Signal Processing and Communication Application Conference (SIU), pp. 133-136, IEEE, 2016.

[19] (n.d.). Retrieved from Hamari Web: https://hamariweb.com/

[20] (n.d.). Retrieved from Hamari Web: https://hamariweb.com/

[21] 1 AI@Meta, "Llama 3 Model Card," 2024. Available: https://github.com/metallama/llama3/blob/main/MODEL_CARD.md

[22] S. Kumar, "Noise Reduction in Audio File Using Spectral Gating and FFT by Python Modules," 2023.

[23] J. B. Kaur, "A review: Audio noise reduction and various techniques," International Journal of Advances in Science Engineering and Technology, 2015.

[24] J. Betker, "Better speech synthesis through scaling," 2023.

[25] J. Kong and K. Jia, "HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis," 2020.