OPEN ACCESS
International Journal of Innovations in Science & Technology
RESEARCH & INNOVATION DIVISION
IJIST

Voice Cloning and Synthesis Using Deep Learning: A Comprehensive Study

Adeel Munir[1], Hammad Nasir[1], Madiha Sher[1], Arbab Masood Ahmad[1],
[1]Department of Computer Systems Engineering, University of Engineering and Technology, Peshawar, Pakistan.
***Correspondence:** adeelmodernite@gmail.com, hammadnasir797@gmail.com, madiha@uetpeshawar.edu.pk , arbabmasood@uetpeshawar.edu.pk,

Citation | Munir. A, Nasir. H, Sher. M, Ahmad. A. M, "Voice Cloning and Synthesis Using Deep Learning: A Comprehensive Study", IJIST, Vol. 07 Issue. 03 pp 2225-2235, September 2025
Received| August 09, 2025 **Revised|** August 29, 2025 **Accepted |** September 02, 2025 **Published|** September 06, 2025

This paper reviews current voice cloning and speech synthesis methods. It focuses on the way that deep learning enhances AI-generated voice synthesis in terms of quality, flexibility, and efficiency. We analyze the top AI models in terms of their significance to virtual assistants, dubbing, and accessibility tools: XTTS_v2, Whisper, and Llama 8B. Voice cloning and TTS efforts in Tortoise are improved by XTTs_v2. Based on the multilingual creative transfer, it has a higher speed and shorter time of a computational process, and generates synthetic speech closer to naturalness. Whisper is a transcription model that goes from an audio waveform to text. It simplifies access to audio data. Llama 8B focuses on user question answering for enhancing AI and human interaction. Other related work includes fastSpeech2 [1], Neural Voice Cloning with few Samples [2], and Deep Learning-Based Expressive Speech Synthesis [3], which also contribute to these advancements. This progress enhances machines' ability to communicate in an emotional and human-like way, leading to more sophisticated technology.

Keywords: Voice Cloning, Speech Synthesis, Deep Learning, Multilingual Zero-shot Multi-Speaker TTS (XTTS), Speaker Adaptation, Cross-Lingual TTS, Whisper, Llama 8B

Introduction:

Speech synthesis has advanced dramatically in recent decades, evolving from rigid, robotic-sounding voices to speech that is highly natural and easy to understand. Earlier approaches to speech synthesis had severe limitations, including monotonous intonation, poor adaptability across different speakers, and the need for enormous amounts of data just to achieve limited performance. Because of these constraints, older systems produced output that was far from lifelike and suitable only for static, non-interactive applications. The field underwent a major transformation with the advent of deep generative models such as WaveNet [1], Tacotron [4], and FastSpeech [5]. These breakthroughs significantly improved speech quality, naturalness, and generation speed, fundamentally reshaping the state of the art. Alongside these innovations, voice cloning techniques also progressed from speaker-dependent approaches to few-shot and zero-shot adaptation methods [6], greatly expanding the scope of voice synthesis applications.

Despite these advances, significant challenges remain. Current systems still struggle with prosody modeling, emotional expressiveness, and consistent quality across different languages. These issues limit the ability to create speech synthesis systems that go beyond intelligibility to being context-aware and culturally inclusive. For instance, although multilingual and cross-lingual synthesis has improved through shared phonetic representations and multilingual datasets, uneven data availability across languages continues to cause disparities in performance. Similarly, while expressive synthesis has made progress in capturing emotions, tone, and conversational nuances, generating voices that are genuinely empathetic and situationally aware remains incomplete.

This work explores these gaps by examining recent developments at the intersections of large-scale modeling, multimodal learning, and creative synthesis. Specifically, we analyze XTTS_v2, Whisper, and Llama 8B, comparing their architectures, training strategies, and applications in high-fidelity voice assistants. By investigating how these systems push forward naturalness, adaptability, and inclusivity, this work highlights the transition of speech synthesis from a narrow technical focus on intelligibility to a broader vision of personalized, contextual, and globally inclusive voice solutions.

Objectives:

The main objectives of this proposed study are:

1. To examine whether XTTS_v2, Whisper, and LLaMA 8B are effective for voice cloning and speech synthesis across multilingual and cross-lingual contexts.
2. To discuss the models in terms of naturalness, adaptability, and training data efficiency using metrics such as MOS, CER, and SECS.
3. To assess how well the models combine with speech recognition (Whisper) and contextual modeling (LLaMA 8B) with cross-lingual voice synthesis (XTTS_v2) in real-time, with effective use of resources.
4. To evaluate how novel TTS pipelines can be applied in real-world contexts, specifically for accessibility, disaster communication, and global AI assistants.
5. To mark out the limitations of currently available audio cloning systems, particularly with reference to emotional expressiveness, prosody modeling, and inclusivity for under-resourced languages, is an important area for future research.

Novelty Statement:

This research advances the field of speech synthesis by comparing three key approaches to virtual transcription and spoken contextual modeling: XTTS_v2, Whisper, and LLaMA-8B. It emphasizes their respective strengths in transcription accuracy, contextual understanding, and cross-lingual voice cloning. In particular, it discusses how XTTS_v2 demonstrates the ability to generate speech that is nearly indistinguishable from human voices while requiring minimal training data, with further extensions into few-shot

and zero-shot adaptation. A clear application is outlined in relation to disaster tweets, where natural language processing (NLP) preprocessing of the tweets using XLNet embeddings linked to expressive, real-time communication of information to the audience confronted with the crisis. The work also provides new perspectives for reconciling tradeoffs of adaptability, emotional tone, emotional engagement, and synthesis efficiency that seek to extend the analysis of synthetic speech away from intelligibility, to focus analysis on inclusive, culturally adaptive, and emotionally engaged synthetic speech.

Literature Review:

Neural Speech Synthesis:

FastSpeech-2 [5] introduced a non-autoregressive architecture that enhanced Tacotron-based model training while significantly improving inference speed, thereby addressing key training challenges. The ability to generate speech more rapidly and stably without autoregressive decoding increases its suitability for real-time applications. Moreover, incorporating speaker conditioning in multi-speaker neural TTS systems has improved speaker generalization, enabling more natural multi-voice synthesis, as demonstrated by Deep Voice-2 [12].

Few-Shot Voice Cloning:

Voice cloning with sparse data has been a central concern of recent work. The Neural Voice Cloning with Few Samples model [13] introduced a speaker adaptation approach capable of generating high-quality speech in a new voice using only a small number of utterances. This few-shot learning approach enables rapid speaker adaptation, reducing the dependence on large training datasets while maintaining the naturalness and intelligibility of synthesized speech.

Expressive Speech Synthesis:

Expressiveness in speech synthesis plays a crucial role in generating natural, emotionally rich voices. Deep Learning-Based Expressive Speech Synthesis [15] emphasized the importance of emotional prosody in neural TTS models, demonstrating how pitch variations, rhythm shifts, and intonation changes contribute to human-like speech. These advancements have paved the way for emotion-aware TTS systems, enhancing user experiences in applications such as virtual assistants and dubbing.

Multilingual and Cross-Lingual Speech Synthesis:

Multilingual and cross-lingual speech synthesis enables voice cloning across different languages while maintaining speaker identity. XTTS_v2 [2] extends traditional TTS models by incorporating cross-lingual text-to-speech capabilities, allowing high-quality voice cloning with minimal training data[8]. Such advancements are especially valuable for applications that demand multilingual support, including global AI assistants and localization services.

Automatic Speech Recognition (ASR) for Speech Synthesis:

The phrase that feeds those TTS pipelines with an input feed provides a perfectly valid transcription, which is crucial for better synthesis output. OpenAI's Whisper [10] represents a state-of-the-art ASR model, offering strong robustness in noisy environments along with multilingual support. By analyzing numerous transcription examples with high accuracy, Whisper generates natural, clear synthetic speech. This is especially needed in processes that use both speech-to-text and text-to-speech

These advancements form the basis of our research. Our goal is to improve speaker adaptability, real-time performance, and multilingual voice synthesis in modern TTS systems.

Methodology:

Traditional methods often produced robotic and unnatural voices, requiring extensive data and manual tuning. Consequently, low-quality voiceovers and automated assistants have often failed to meet user expectations, with customer satisfaction reported at only 65% (2023 Voice Technology Report) [11].

The performance metrics from recent models, as shown in Table 1, further highlight the need for further refinement in both voice cloning accuracy and naturalness. Metrics such as Character Error Rate (CER), UTMOS (perceived naturalness), and SECS (speaker similarity) provide a quantitative basis for assessing system performance relative to human speech. Notably, models such as HierSpeech++ and StyleTTS 2 achieved high UTMOS scores, yet CER and SECS revealed persistent gaps in clarity and speaker resemblance.

Table 1. CER, UTMOS, and SECS for English Language Speech Synthesis Models

Model	Hours	CER (↓)	UTMOS (↑)	SECS (↑)
Tortoise	4yk	1.0y34	4.0883 ± 0.31	0.54y2
StyleTTS 2	245	0.578y	4.4250 ± 0.07	0.4728
HierSpeech++	2.7k	0.7741	4.457 ± 0.05	0.5530
YourTTS (Original)	474 / 28y (en)	2.8735	3.5034 ± 0.2y	0.4521
YourTTS (Exp. 1)	245	1.0y1	4.102 ± 0.25	0.7120
XTTS (Exp. 3)	27k / 14k (en)	0.5425	4.007 ± 0.25	0.5423

Building on these insights, our project sought to improve speech synthesis using modern techniques, enhancing voice cloning via small data footprints as well as cross-lingual expressivity. In particular, we utilized recent natural language processing (NLP) pipelines with advanced text-to-speech (TTS) models to explore communication situations in disasters. We collected and employed tweets as input data to mimic real-time, brief text communication that was synthesized into natural and expressive speech. In addressing these issues, we hoped to contribute to the goal of creating AI-generated voices that are more natural and context-sensitive in the areas of accessibility, content generation, and crisis communication.

Text Stream:

We explored the opportunity Twitter represents as an important source of information during and following disasters. We identified a series of relevant tweets using the Twitter API and focused on both tweets containing hashtags specific to disasters and geotagged tweets. In total, we collected 5,313 tweets to use to build a candidate classification model for future use.

Once completed, the tweets were preprocessed. This included cleaning up the data and removing unnecessary tweets, such as advertisements, duplicates, and posts in languages other than English. The remaining tweets were organized into a specific format. This ensured that the tweets represented disaster-related communication and were suitable for use in training of NLP and XTTS_v2 model [2] inputs.

Text Pre-Processing:

In the field of natural language processing, preprocessing plays a vital role in converting raw data into a form that is more suitable for training a model [12]. In our case, we used a multi-step pipeline where we:

1. Remove URLs, hashtags, user mentions, and excess whitespace.
2. Standardize Unicode text to eliminate inconsistencies.
3. Filter out irrelevant tweets (i.e., off-topic, spam, etc.).

After cleaning, the textual dataset yielded a more consistent format for representing the discourse surrounding disasters and enhanced the rigor of the modeling in later steps. For our work in NLP, we chose the XL-Net architecture. Compared to earlier models such as BERT, it was better at capturing long-range context and understanding language in both directions. It was overall effective for short, noisy, and context-dependent texts like tweets.

Text Normalization: Normalization of the text, the next stage in natural language analysis, standardizes the text information for analysis. It includes a series of steps that involve removing punctuation, converting all text to lowercase, and eliminating special characters.

After the text is pre-processed, normalization techniques are applied to ensure further standardization. These include:

1. All text converted to lower case.
2. Removing punctuation and special characters.
3. Expanding common contractions and abbreviations (i.e., "can't" → "cannot").

These steps improved semantic consistency and maintained compatibility with the XLNet embeddings. The normalization process also enabled token-level analysis, allowing the model to better capture nuanced meanings often present in short, disaster-related messages.

Tokenization:

Tokenization was used to break text into meaningful units. We chose subword-level tokenization because disaster-related tweets often included rare terms, abbreviations, and specific jargon. Unlike word-level tokenization, which has trouble with unknown words, subword segmentation offered flexibility in managing previously unseen terms while keeping the meaning intact. This ensured that both common and unique linguistic elements were well represented for model training.

Model Training and Integration with Speech Synthesis:

The XLNet model received the tokenized dataset to extract contextual embeddings. These embeddings were connected to speech synthesis pipelines using XTTS_v2. We chose XTTS_v2 because it can work with minimal training data [13] to achieve cross-lingual voice cloning, which helps address the challenges of disaster communication that often requires multilingual support.

By combining text analytics with modern TTS architectures, the system aimed to unite linguistic meaning with vocal naturalness. The authors also worked to reduce the training data needed for voice cloning while ensuring cross-lingual recognition and emotional expression. This system allowed disaster-relief tweets to be transformed into natural, clear, and empathetic outputs in real time.

Models:

XTTS_V2 Model:

XTTS_v2, a model introduced by Coqui AI, is one of the big steps towards innovative text-to-speech (TTS) systems that can perform voice cloning with a very small amount of training data [8]. This model uses a transformer-based encoder-decoder architecture [2][14] approach, which enables it to generate speech fast and effectively with high quality. It also performs well with different speakers through fine-tuning, which equips it to mimic a specific speaker's vocal traits very closely. One of its key advantages is its ability to produce high-quality speech while maintaining emotional tone and expressiveness, making the synthesized voice sound natural and real. These features make XTTS_v2 a useful voice synthesis tool.

Why XTTS:

User preference scores Table 2 offers qualitative judgments of XTTS relative to HierSpeech++ and Mega-TTS. These evaluations are shown by the Comparative Mean Opinion Score (CMOS), which reflects overall user preference. The Speaker Mean Opinion Score (SMOS) captures how similar users find the speakers.

Table 2. Comparison of CMOS and SMOS scores between XTTS and other models

Comparison	CMOS (↑)	SMOS (↑)
XTTS vs HierSpeech++	0.41 ± 0.25	-0.31 ± 0.35
XTTS vs Mega-TTS 2	0.y2 ± 0.22	-0.3y ± 0.38

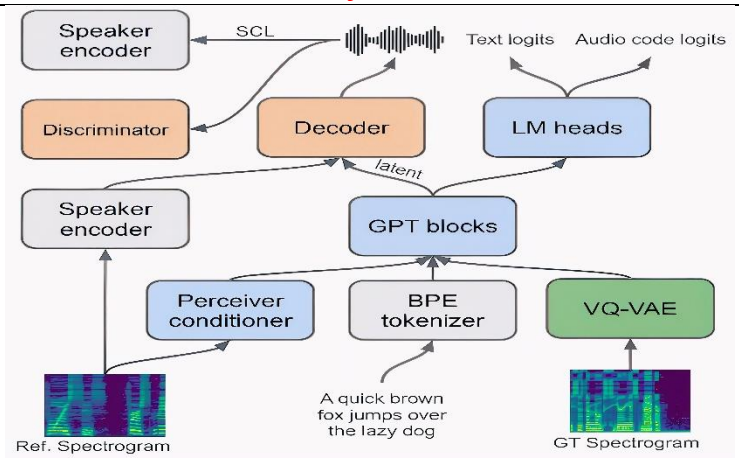

Figure 1. XTTS training architecture overview.

Interpretation of Results:

The results in Table 2 highlight the strengths and weaknesses of XTTS_v2, which are:

User Preference (CMOS):

Listeners consistently rated XTTS higher in naturalness and overall user experience compared to HierSpeech++ and Mega-TTS. The positive CMOS values of 0.41 and 0.72 back this up, as evident from Figure 2.

This suggests that XTTS creates speech that sounds more natural. It matches previous studies showing it performs well even with limited data, surpassing earlier systems like WaveNet [1], Tacotron [4], and FastSpeech [5].

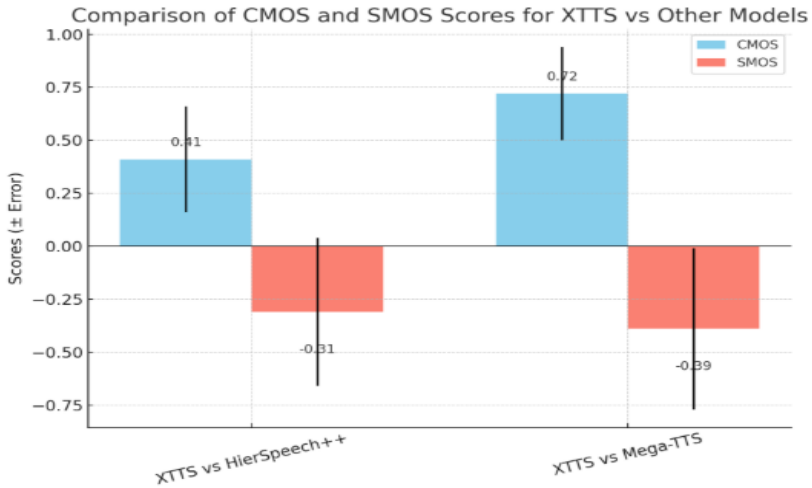

Figure 2. Comparison of CMOS Scores for XTTS vs Other Models

Speaker Likeness (SMOS):

Despite its high naturalness shown in Figure 3, XTTS received a negative score on SMOS, at -0.31 compared to HierSpeech++ and -0.39 compared to Mega-TTS.

This shows that listeners felt there was a weaker connection between the voices produced and the target speaker identities. It points out a trade-off between naturalness and speaker similarity.

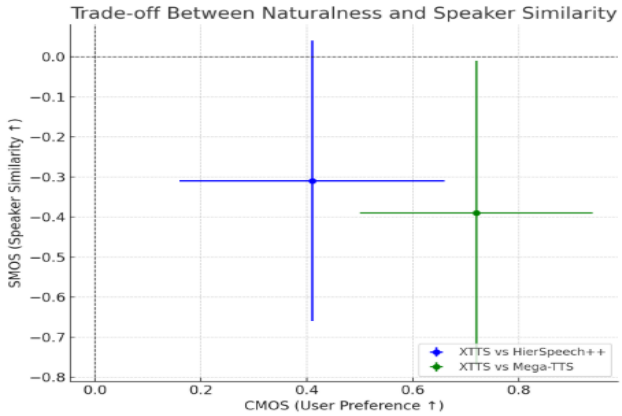

Figure 3. SMOS Score showing Between Naturalness and Speaker Similarity

Overall Implications:

XTTS is very good at producing expressive, natural-sounding speech, making it a strong option for applications that prioritize user experience and emotional tone.

However, it needs to improve in modeling speaker-specific features, like consistent prosody and unique vocal traits, to better maintain speaker identity.

These findings highlight a key challenge in speech synthesis: finding the right balance between perceived quality and accurate speaker identity.

XTTS and Whisper Integration:

XTTS has gained significant attention due to its advanced multilingual voice cloning capabilities. Whereas earlier models tend to be limited to a single language or need a lot of speaker adaptation, XTTS can synthesize 17 different languages while keeping the speaker's identity intact.

Whisper Model:

The speech recognition and transcription model Whisper works well in noisy environments [10] because OpenAI built a strong product. The system turns spoken words into accurate text transcriptions for high-quality use in other applications. Text-to-speech systems gain from Whisper's pre-processing since it helps improve accuracy with properly formatted input text. The model shows a strong understanding of multiple languages because it detects and transcribes them effectively [10]. This combination of features makes Whisper an important part of modern speech technology and improves application efficiency through voice-based interfaces.

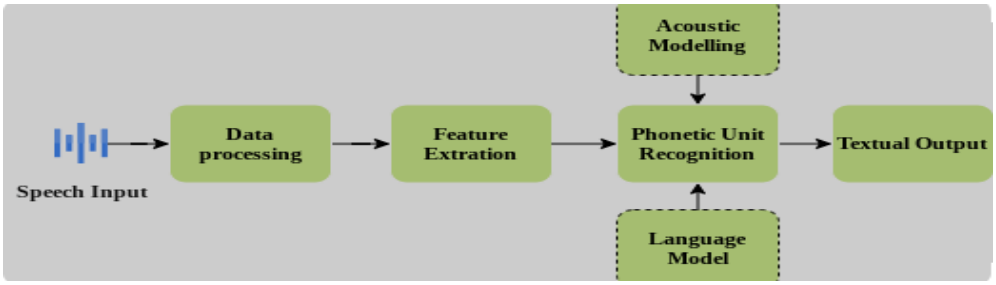

Figure 2. Pipeline integration of Whisper speech-to-text Generation

Training LLaMA 8B as a large-scale language system gives it primacy in creating new textual outputs and generating insights that maintain emotional sensitivity. The model helps understand text context in detail to execute precise, relevant language operations. The alignment between speech synthesis models at the phoneme level receives improvements through the system to achieve better synchronization between text and generated speech. With its optimized speech-to-text functionality, LLaMA 8B enhances the natural and dynamic voice output capability with its applications aligned with speech models [3].

Experimental Setup and Results:

An analysis was conducted after testing and observing the performance of XTTS_v2, Whisper, and LLaMA 8B using different key metrics that measured their efficiency in speech recognition and synthesis. To assess the natural quality of synthesized speech, the Mean Opinion Score (MOS) was chosen as the tool for subjective evaluation [12]. The evaluation of synthesis speed identified which model could produce speech output the fastest in real-time. The assessment of speaker adaptability looked at each model's minimum training requirement for voice cloning and its ability to replicate individual voice traits [13]. Under real-world conditions, these tests provide a complete view of both the strengths and limitations of each model.

Table 3. Evaluating Models: MOS Scores and Training Data

Model	MOS Score	Training Data Required
XTTS_v2	4.5	Low
Whisper	4.4	Medium
LLaMA 8B	4.5	High

Analysis of MOS Score:

XTTS_v2 and LLaMA 8B both achieved a MOS of 4.5, as shown in Table 3, which can be visualized in Figure 4, showing they closely resemble human speech and text generations, respectively. This sets a new record for models like WaveNet [1] and Tacotron [4], which have relatively low Mean Optimum Score (MOS).

Whisper also did well with an overall MOS of 4.4, clearly showing high-quality speech transcription abilities. This shows that all three systems together produced speech outputs rated as clear and natural. The near-human scores further support the potential of these systems for high-quality voice applications.

Training Data Requirements and Speaker Adaptability:

A key difference among the models discussed is their training data needs, evident from Figure 4 and Table 3 as well. XTTS_ v2 required less data and showed a high level of flexibility, making it suitable for few-shot or zero-shot voice cloning applications. This contrast is significant when compared to earlier systems like FastSpeech [5], which needed a lot of data and struggled with speaker generalization reliability. Whisper fell into the moderate data category and demonstrated good predictive transcribing capability. This helped improve synthesis quality by reducing transcription errors in the input text. LLaMA 8B required much more data for training but offered better contextual modeling and more refined generation abilities [3]. These findings highlight a trade-off between flexibility and richer contextual modeling in speech synthesis design.

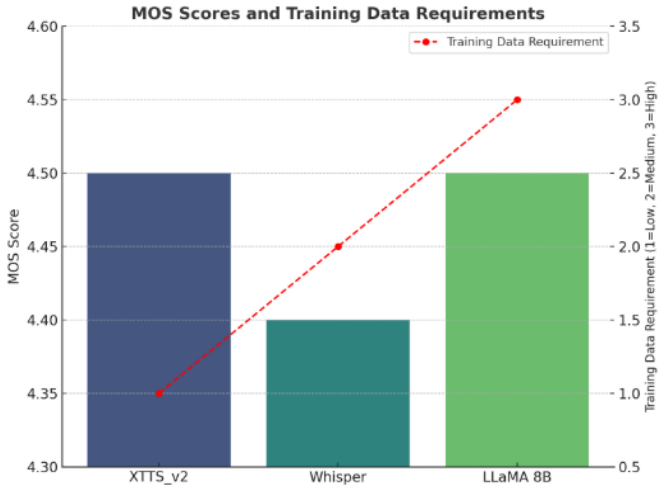

Figure 4. MOS scores and training data needs of XTTS_v2, Whisper, and LLaMA 8B.

Summary of Findings:

Collectively, the results show that:

XTTS_v2 provides the best mix of high MOS scores, low training data requirements [13], and real-time readiness; it is suitable for scalable voice cloning applications.

Whisper has strong transcription capabilities and decent speech synthesis quality. This makes it especially valuable for applications where the quality of input text is a key factor.

LLaMA 8B has better contextual awareness and naturalness than XTTS_v2 and Whisper, but it requires more computational resources and training data [3].

These results support the transition identified in the prior research: from models constrained by intelligibility and data demands (i.e., WaveNet, Tacotron, FastSpeech) [1][4][5] to systems that are more adaptive, efficient, and context-aware. The observed trade-off suggests that future work can consider hybrid pipeline systems by combining Whisper transcriptions for accuracy, XTTS_v2 voice cloning for efficiency, and LLaMA 8B for richer contextual modeling [3], to transition toward the next generation of globally-inclusive and emotionally-aware speech synthesis systems.

Discussion:

Advantages of XTTS_v2 and Whisper:

The integration of XTTS_v2, Whisper, and LLaMA 8B demonstrated significant advantages over traditional and earlier deep learning–based speech synthesis methods. XTTS_v2 was particularly effective in producing high-quality voice cloning results with minimal training data [13] requirements, making it an economical and scalable solution. In contrast to earlier models like WaveNet [1], Tacotron [4], and FastSpeech [5], which needed large datasets and extensive training to work well with different speakers, XTTS_v2 performed strongly even in few-shot and zero-shot situations.

Likewise, Whisper provided reliable transcription, ensuring that the text inputs used for creating speech were of high quality. This solved a problem seen in earlier systems, where transcription errors led to poor speech output. The language modeling abilities of LLaMA 8B further helped in producing speech that was more aware of context and sounded more natural, going beyond the simpler sequence-to-sequence language models used before. Together, these models moved speech synthesis beyond just making speech understandable, toward systems that are flexible, efficient, and responsive to context.

Challenges and Future Improvements:

Despite these advances, challenges remain that reflect broader gaps found in earlier work. While XTTS_v2 provides data efficiency, achieving consistent emotional expressiveness is still a problem. Earlier models like Tacotron and FastSpeech had difficulty with prosody and emotion modeling [8]. Although newer methods have brought some improvements, generating speech with natural emotional depth is still limited. This indicates that future efforts should focus on prosodic variation and emotion-aware modeling to boost expressiveness and user engagement [9][11].

Another challenge is cross-lingual synthesis. Although XTTS_v2 and similar models have progressed in multilingual adaptation, performance is still inconsistent for under-resourced languages. Previous studies that used shared phonetic representations and multilingual corpora [14][15] showed promising directions, but achieving fair quality across languages still needs larger, more diverse, and well-curated datasets.

Finally, while LLaMA 8B improves contextual modeling, its integration with multimodal cues like gestures, dialogue context, and situational awareness is still not well explored. Advancing toward truly human-like voice assistants will require not only technical refinements in modeling but also broader consideration of cultural, linguistic, and emotional inclusivity.

Deep Learning Approaches in Speech Synthesis:

Sequence-to-Sequence Models:

Tacotron and Tacotron 2 have redefined speech synthesis by directly mapping text sequences to spectrograms without passing through traditional intermediate representations. With a WaveNet vocoder, Tacotron 2 combines a recurrent neural network (RNN)- based text encoder for high-quality, natural speech. Nevertheless, these models tend to be quite fragile and require considerable post-processing.

Transformer-Based Models:

Transformer architectures have found their place in speech synthesis with their parallelism and scalability. fastSpeech and its successors use transformer encoders and decoders to achieve fast and stable non-autoregressive synthesis. With speech data, these architectures excel at modeling long-range dependencies and capturing intricate patterns.

Neural Vocoders:

WaveNet, Parallel WaveGAN, and Hifi-GAN have altogether transformed the wave generation process. These models generate high-fidelity speech directly from spectrogram inputs, significantly improving naturalness and clarity. Neural vocoding has been instrumental in bridging the gap between synthetic and human-like audio.

Conclusion and Future Work:

This study demonstrated the effectiveness of XTTS_v2, Whisper, and LLaMA 8B in achieving high-fidelity voice cloning and synthesis. The results highlight their capabilities in generating natural-sounding speech with accurate transcription and contextual understanding. However, several areas for future research remain. One key direction is expanding multilingual training to improve speaker generalization across diverse languages and accents. Additionally, enhancing real-time TTS performance is essential for applications such as AI assistants and automated dubbing, where low-latency synthesis is crucial. Furthermore, integrating advanced tone and emotion recognition can refine speech expressiveness [5][8], making AI-generated voices more natural and engaging. Addressing these areas will further advance the field of AI-driven speech synthesis.

Acknowledgments:

We are profoundly grateful to all those who helped us in this research project, especially the faculty and staff of the University of Engineering and Technology, Peshawar, Pakistan.

The manuscript has not been published or submitted to other journals previously.

Author's Contribution:

It is acknowledged that all authors have contributed significantly and that all authors agree with the content of the manuscript.

Conflict of Interest:

There exists no conflict of interest for publishing this manuscript in IJIST.

References:

[1] K. S. Aaron van den Oord, Sander Dieleman, Heiga Zen, "WaveNet: A Generative Model for Raw Audio," *arXiv:1609.03499*, 2016, doi: https://doi.org/10.48550/arXiv.1609.03499.

[2] L. H. Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, "XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model," *arXiv:2406.04904*, 2024, doi: https://doi.org/10.48550/arXiv.2406.04904.

[3] X. M. Hugo Touvron, Thibaut Lavril, Gautier Izacard, "LLaMA: Open and Efficient Foundation Language Models," *arXiv:2302.13971*, 2023, doi: https://doi.org/10.48550/arXiv.2302.13971.

[4] Y. Wang *et al.*, "Tacotron: Towards End-to-End Speech Synthesis," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2017-August, pp. 4006–4010, 2017,

doi: 10.21437/INTERSPEECH.2017-1452.

[5] T.-Y. L. Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," *arXiv:2006.04558*, 2020, doi: https://doi.org/10.48550/arXiv.2006.04558.

[6] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 7962–7966, Oct. 2013, doi: 10.1109/ICASSP.2013.6639215.

[7] Y. W. Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis," *Adv. Neural Inf. Process. Syst.*, vol. 31, pp. 4485–4495, 2018, doi: https://doi.org/10.48550/arXiv.1806.04558.

[8] J. Latorre *et al.*, "Effect of Data Reduction on Sequence-to-sequence Neural TTS," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2019-May, pp. 7075–7079, May 2019, doi: 10.1109/ICASSP.2019.8682168.

[9] N. S. Suparna De, Ionut Bostan, "Making Social Platforms Accessible: Emotion-Aware Speech Generation with Integrated Text Analysis," *16th Int. Conf. Adv. Soc. Networks Anal. Min. -ASONAM-2024*, 2024, doi: https://doi.org/10.48550/arXiv.2410.19199.

[10] I. S. Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, "Robust Speech Recognition via Large-Scale Weak Supervision," *arXiv:2212.04356*, 2022, doi: https://doi.org/10.48550/arXiv.2212.04356.

[11] Y. Z. Zuen Cen, "Investigating the Impact of AI-Driven Voice Assistants on User Productivity and Satisfaction in Smart Homes," *J. Econ. Theory Bus. Manag.*, vol. 1, no. 6, pp. 8–14, 2024, doi: 10.70393/6a6574626d.323333.

[12] Y. Z. Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, "Deep Voice 2: Multi-Speaker Neural Text-to-Speech," *arXiv:1705.08947*, 2017, doi: https://doi.org/10.48550/arXiv.1705.08947.

[13] Y. Z. Sercan O. Arik, Jitong Chen, Kainan Peng, Wei Ping, "Neural Voice Cloning with a Few Samples," *arXiv:1802.06006*, 2018, doi: https://doi.org/10.48550/arXiv.1802.06006.

[14] I. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, p. 30, 2017.

[15] H. B. Chandran, "Deep learning-based expressive speech synthesis," *EURASIP J. Audio, Speech, Music Process.*, 2024, doi: https://doi.org/10.1186/s13636-024-00329-7.