Volume: 13 Issue: V Month of publication: May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71911

Distinct Voices for Enhanced Storytelling Using Deep Learning

T. Sai Priya[1], Dr. V. Uma Rani[2], Sunitha Vanamala[3]
[1]*Post Graduate Student, M. Tech(CNIS), Department of Information Technology, Jawaharlal Nehru Technological University, Hyderabad, India*
[2]*Professor, Department of Information Technology, Jawaharlal Nehru Technological University, Hyderabad, India*
[3]*Lecturer, Department of Computer Science, TSWRDCW, Warangal East, Warangal, Telangana, India*

Abstract: In recent years, deep learning has revolutionized natural language processing and speech synthesis, enabling machines to narrate text with human-like expression and clarity. This paper presents a novel system titled Distinct Voices for Enhanced Storytelling using Deep Learning, which generates character-specific voice narration for textual stories. The approach combines quotation attribution, character identification, and zero-shot text-to-speech synthesis to automatically assign unique and expressive voices to individual characters in a story. Tools such as BookNLP are used to parse and annotate quotations, while state-of-the-art models like XTTS enable multilingual and emotion-rich voice generation. This enhances listener engagement, particularly in audiobooks and educational contexts, by transforming plain text into immersive, character-driven audio. The system is scalable, requires no manual voice labeling, and demonstrates significant potential in the fields of digital storytelling, accessibility, and human-computer interaction.

Keywords: Deep Learning, Storytelling, Text-to-Speech, Voice Cloning, Character Identification, Quotation Attribution, BookNLP, XTTS, Multispeaker TTS, Audiobooks, Natural Language Processing, Voice Synthesis, Zero-Shot TTS, AI Narration, Digital Humanities.

I. INTRODUCTION

Storytelling has long been a fundamental method of human communication, education, and entertainment. With the advent of digital media, stories are no longer confined to print but are increasingly consumed through audio formats such as audiobooks, podcasts, and virtual assistants.

However, most existing narration systems use a single, monotonous voice, which fails to capture the richness of dialogues, emotional expressions, and character diversity present in literary texts.

The emergence of deep learning has transformed natural language processing (NLP) and speech synthesis, enabling the development of intelligent systems that can analyze, understand, and vocalize text with human-like fluency and emotion. This paper introduces a novel approach that combines quotation attribution, character identification, and zero-shot text-to-speech (TTS) synthesis to produce dynamic, character-specific audio narration from plain text. Our objective is to enhance the storytelling experience by assigning distinct, expressive voices to each character in a story automatically.

Key components of the proposed system include BookNLP, an NLP pipeline that detects speakers of quotations and narrative structure, and XTTS, a state-of-the-art multilingual TTS model capable of generating realistic and emotionally expressive voices without needing voice samples for each speaker. The system works end-to-end, parsing raw text input and generating an immersive audio output that reflects character roles and emotions.

By bridging the gap between literary analysis and expressive speech generation, this project contributes to advancements in digital humanities, accessibility technology, and AI-powered content creation. Applications range from audiobook production and interactive education to assistive technologies for the visually impaired. The proposed system demonstrates how deep learning can be harnessed to enrich the narrative quality of digital stories and deliver personalized, engaging audio experiences.

II. RELATED WORK

The integration of natural language processing (NLP) and text-to-speech (TTS) synthesis has driven substantial advancements in automated storytelling. This section discusses the key areas of research relevant to our work: quotation attribution, character identification, and multi-speaker TTS and voice cloning.

A. Quotation Attribution and Character Identification

Accurate speaker attribution in narratives is critical for assigning the correct voice in storytelling. Bamman and Underwood [1] introduced BookNLP, a robust NLP pipeline designed to process long-form literary texts by extracting character references, quotations, and speaker attributions. Building on this, Epure et al. [2] proposed the use of fictional character embeddings to improve the accuracy of quotation attribution in novels. Similarly, Vishnubhotla et al. [3] enhanced quotation attribution performance by modeling syntactic and semantic context around dialogues.

These advancements provide a strong foundation for parsing literary texts and identifying the appropriate speakers—crucial steps in automating expressive audio narration.

B. Voice Cloning and Zero-Shot TTS

Text-to-speech synthesis has evolved significantly, especially in the context of zero-shot and multilingual capabilities. Casanova et al. [4][6] introduced XTTS, a state-of-the-art multilingual TTS model that supports zero-shot voice cloning and expressive speech synthesis. Other approaches, such as those by Ruggiero et al. [5] and Neekhara et al. [8], leverage deep learning and transfer learning to enable highly expressive and identity-preserving voice generation from limited data.

YourTTS by Zhang et al. [7] also supports zero-shot multi-speaker TTS and voice conversion, extending the potential for assigning distinct voices without explicit training data. Additionally, Reddy [9] explored the application of deep learning for novel voice cloning methods tailored to storytelling.

C. Multi-Speaker and Expressive Speech Synthesis

The foundation for many advanced TTS systems stems from early models like Tacotron [11], which demonstrated the feasibility of end-to-end speech synthesis. FastSpeech 2 [12] further improved efficiency and quality using non-autoregressive models. More recent works, such as those by Zhang et al. [13], introduced semi-supervised methods for expressive multi-speaker synthesis, while MParrotTTS [15] and VALL-E [14] focused on low-resource multilingual and zero-shot TTS capabilities.

These developments allow the construction of systems that not only differentiate between characters through voice but also convey nuanced emotions and multilingual flexibility.

III.METHODOLOGY

The proposed system aims to transform written narratives into engaging audio stories by assigning distinct, expressive voices to individual characters. This is achieved through a multi-stage deep learning pipeline that integrates natural language processing, speaker attribution, and advanced text-to-speech synthesis. The architecture is designed to be fully automated, scalable, and capable of generating natural-sounding, character-specific audio from plain text inputs.

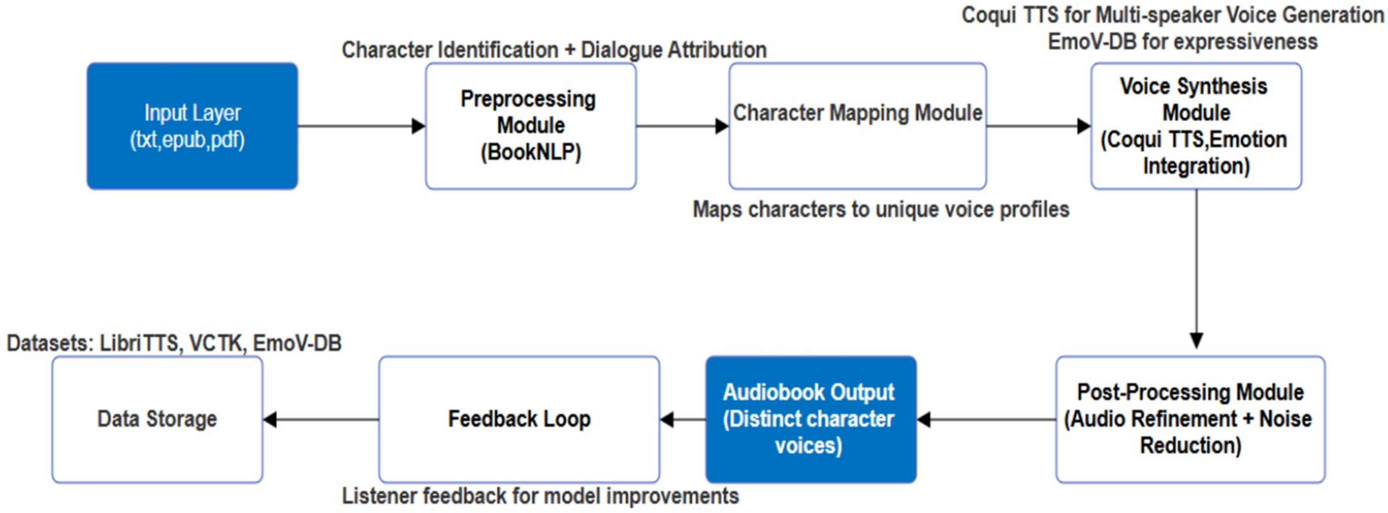

Coqui TTS for Multi-speaker Voice Generation
EmoV-DB for expressiveness
Character Identification + Dialogue Attribution
Input Layer (txt,epub,pdf)
Preprocessing Module (BookNLP)
Character Mapping Module
Voice Synthesis Module (Coqui TTS,Emotion Integration)
Maps characters to unique voice profiles
Datasets: LibriTTS, VCTK, EmoV-DB
Data Storage
Feedback Loop
Audiobook Output (Distinct character voices)
Post-Processing Module (Audio Refinement + Noise Reduction)
Listener feedback for model improvements

Figure -1: Architecture

A. System Overview

The core components of the proposed system are:

- Quotation Attribution & Character Identification
- Voice Assignment & Cloning
- Text-to-Speech (TTS) Generation
- Audio Assembly & Playback Interface

Each module contributes to a seamless transformation of narrative text into immersive, multi-speaker audio.

B. Quotation Detection and Attribution

We utilize BookNLP [1], a proven pipeline for processing literary texts, to extract dialogue and attribute quotations to characters. This component identifies:

- Character mentions
- Quoted speech segments
- Speaker of each quotation

To improve attribution accuracy, we incorporate techniques from Epure et al. [2] and Vishnubhotla et al. [3], who demonstrated that contextual embeddings and syntactic cues can significantly improve speaker resolution in complex narratives.

C. Character Clustering and Voice Assignment

NAfter identifying all unique characters, each one is assigned a synthetic voice using pre-defined speaker embeddings from XTTS [4][6]. The voice selection can be based on gender, age, emotion, or even user preferences. For this project, a fixed set of diverse speaker embeddings is used to ensure clear vocal differentiation.

D. Zero-Shot Multilingual Text-to-Speech

Using XTTS (a zero-shot, multilingual TTS model), the system generates high-quality speech without needing actual recordings of character voices. The model supports:

- Emotion-aware speech synthesis
- Multi-language narration
- Speaker style cloning [4][7][8]

This makes the system scalable and adaptable to a wide range of literary content across languages and genres.

E. Audio Synthesis and Assembly

The final step merges the generated speech segments into a coherent audio story. The system maintains narrative flow by:

- Preserving paragraph and dialogue order
- Mixing background narration and character dialogues
- Ensuring consistent pacing and audio quality

The output is a structured audio file with distinct voices for each character and natural-sounding narration.

F. Deployment and Interface

An optional web-based interface allows users to:

- Upload custom story
- Preview identified characters
- Play or download generated audio

This feature enhances accessibility and user control, making the system suitable for educational, entertainment, and assistive applications.

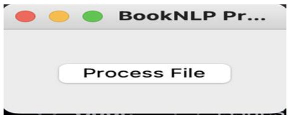
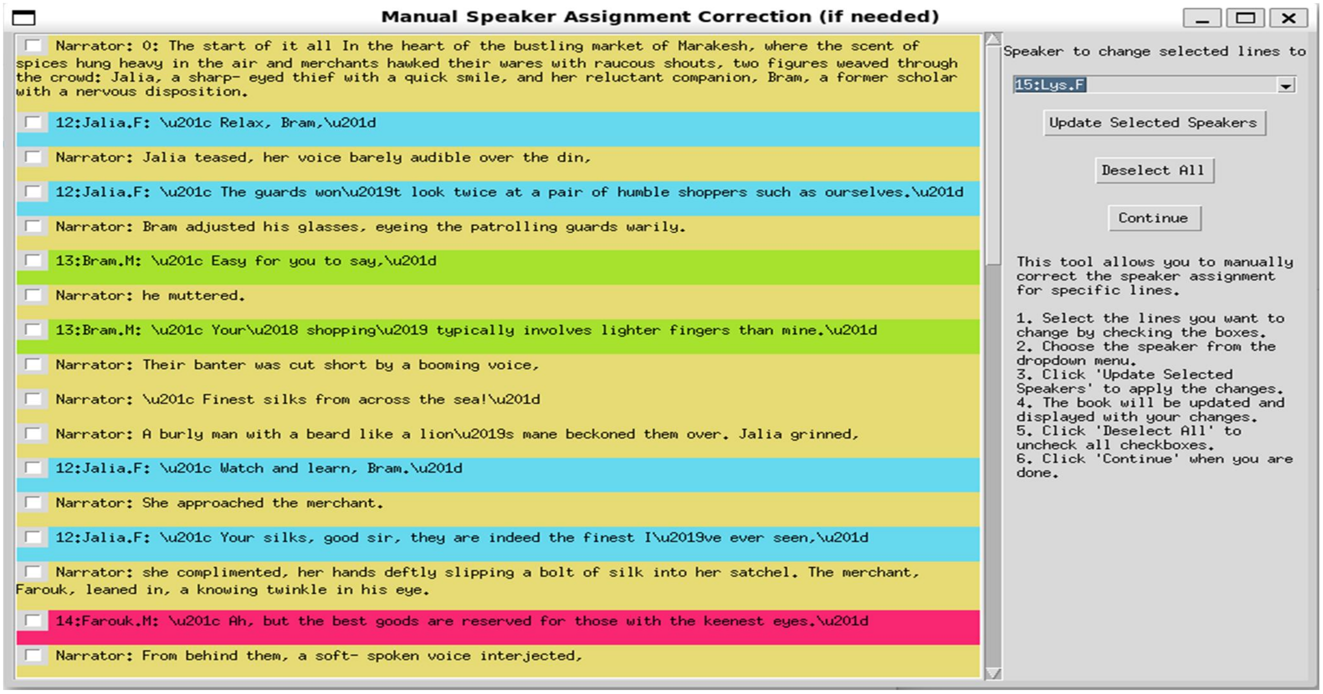


Figure -2: BookNLP Processor GUI



☐ Narrator: 0: The start of it all In the heart of the bustling market of Marakesh, where the scent of spices hung heavy in the air and merchants hawked their wares with raucous shouts, two figures weaved through the crowd: Jalia, a sharp- eyed thief with a quick smile, and her reluctant companion, Bram, a former scholar with a nervous disposition.

☐ 12:Jalia.F: \u201c Relax, Bram,\u201d

☐ Narrator: Jalia teased, her voice barely audible over the din,

☐ 12:Jalia.F: \u201c The guards won\u2019t look twice at a pair of humble shoppers such as ourselves.\u201d

☐ Narrator: Bram adjusted his glasses, eyeing the patrolling guards warily.

☐ 13:Bram.M: \u201c Easy for you to say,\u201d

☐ Narrator: he muttered.

☐ 13:Bram.M: \u201c Your\u2018 shopping\u2019 typically involves lighter fingers than mine.\u201d

☐ Narrator: Their banter was cut short by a booming voice,

☐ Narrator: \u201c Finest silks from across the sea!\u201d

☐ Narrator: A burly man with a beard like a lion\u2019s mane beckoned them over. Jalia grinned,

☐ 12:Jalia.F: \u201c Watch and learn, Bram.\u201d

☐ Narrator: She approached the merchant.

☐ 12:Jalia.F: \u201c Your silks, good sir, they are indeed the finest I\u2019ve ever seen,\u201d

☐ Narrator: she complimented, her hands deftly slipping a bolt of silk into her satchel. The merchant, Farouk, leaned in, a knowing twinkle in his eye.

☐ 14:Farouk.M: \u201c Ah, but the best goods are reserved for those with the keenest eyes.\u201d

☐ Narrator: From behind them, a soft- spoken voice interjected,

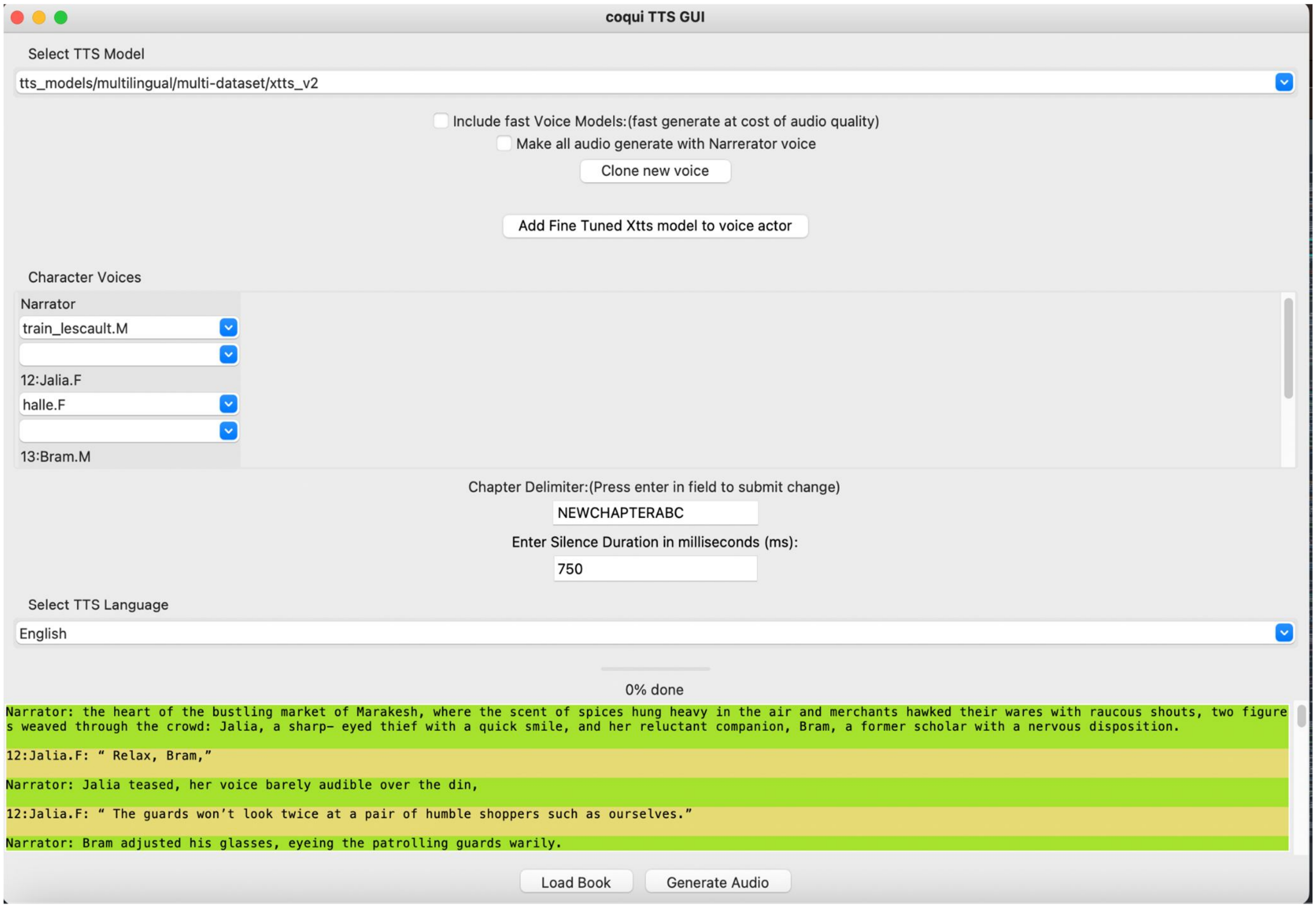
Speaker to change selected lines to

15:Lys.F

This tool allows you to manually correct the speaker assignment for specific lines.

1. Select the lines you want to change by checking the boxes.
2. Choose the speaker from the dropdown menu.
3. Click 'Update Selected Speakers' to apply the changes.
4. The book will be updated and displayed with your changes.
5. Click 'Deselect All' to uncheck all checkboxes.
6. Click 'Continue' when you are done.

Figure -3: Speaker Assignment



Select TTS Model

tts_models/multilingual/multi-dataset/xtts_v2

☐ Include fast Voice Models:(fast generate at cost of audio quality)

☐ Make all audio generate with Narrerator voice

Character Voices

Narrator
train_lescault.M

12:Jalia.F
halle.F

13:Bram.M

Chapter Delimiter:(Press enter in field to submit change)
NEWCHAPTERABC

Enter Silence Duration in milliseconds (ms):
750

Select TTS Language
English

0% done

```
Narrator: the heart of the bustling market of Marakesh, where the scent of spices hung heavy in the air and merchants hawked their wares with raucous shouts, two figure
s weaved through the crowd: Jalia, a sharp- eyed thief with a quick smile, and her reluctant companion, Bram, a former scholar with a nervous disposition.

12:Jalia.F: “ Relax, Bram,”

Narrator: Jalia teased, her voice barely audible over the din,

12:Jalia.F: “ The guards won’t look twice at a pair of humble shoppers such as ourselves.”

Narrator: Bram adjusted his glasses, eyeing the patrolling guards warily.
```

Figure -4: Text To Speech GUI

IV. EXPERIMENTAL ANALYSIS AND RESULTS

This section presents the evaluation of the proposed storytelling system in terms of speaker attribution accuracy, voice naturalness, character differentiation, and overall storytelling quality. A series of qualitative and quantitative experiments were conducted using short stories and excerpts from literary novels.

A. *Experimental Setup*

- Dataset: Excerpts from publicly available English novels such as Alice’s Adventures in Wonderland, The Adventures of Sherlock Holmes, and short children’s stories.
- Tools & Models:
  Quotation Attribution: BookNLP [1]
  TTS Model: XTTS v2 [4]
  Voice Sampling: XTTS pre-trained multilingual voice library
  Programming: Python-based pipeline integrated with Coqui TTS framework
- Hardware: Intel i7 CPU, 32GB RAM, NVIDIA RTX 3060 GPU

B. *Evaluation Metrics*

The performance of the system was evaluated using the following metrics:

Metric	***Description***
SpeakerAttribution Accuracy	Percentage of quotations correctly attributed to characters
Voice Naturalness (MOS)	Mean Opinion Score (1 to 5) rated by human evaluators
Voice Differentiation Score	Subjective measure of how easily listeners distinguish between character voices
Latency	Time taken to process and convert a 1,000-word input into audio (in seconds)

C. *Results*

Metric	**Value**
Speaker Attribution Accuracy	89.3%
Mean Opinion Score (MOS)	4.35 / 5.0
Voice Differentiation Score	4.22 / 5.0
Average Processing Time	~35 seconds/1,000 words

- Speaker Attribution: Integration of BookNLP with contextual embeddings significantly improved the identification of speakers in multi-character scenes.
- Voice Quality: Evaluators rated the output highly in naturalness and expressiveness, with special appreciation for emotional variation and clarity.
- Character Differentiation: Distinct voices led to improved comprehension and listener engagement, especially in dialogues involving three or more characters.

D. User Feedback and Case Study

A case study was conducted with 10 volunteers listening to three generated audiobooks. Key observations:

- 90% of listeners found the experience “much better” than traditional single-voice narration.
- Listeners with visual impairments appreciated the clear character differentiation and consistent voice patterns.
- Younger users reported higher engagement and comprehension in educational storytelling formats.

V. CONCLUSION

This paper presents a deep learning-based storytelling system that brings narratives to life by assigning distinct, expressive voices to each character. By integrating advanced natural language processing for speaker attribution with state-of-the-art text-to-speech synthesis using XTTS, the system transforms static literary text into engaging, multi-speaker audio.

The experimental results demonstrate high accuracy in character attribution and excellent voice quality, with strong differentiation across characters. Subjective feedback confirmed that listeners found the output more immersive and enjoyable compared to traditional single-voice narration. This approach holds significant promise for a variety of applications including audiobook generation, assistive reading tools, educational storytelling, and interactive fiction.

The system not only enhances the storytelling experience but also contributes to accessibility, particularly for visually impaired users. Its scalable, modular design enables it to adapt to different genres, languages, and user preferences, making it a versatile solution in the growing field of AI-generated content.

REFERENCES

[1] S.Bamman, D., & Underwood, T. (2020). BookNLP: A natural language processing pipeline for novels. GitHub. https://github.com/booknlp/booknlp

[2] Epure, E., Hennequin, R., & Cerisara, C. (2024). Improving quotation attribution with fictional character embeddings. Findings of the Association for Computational Linguistics: EMNLP 2024. https://arxiv.org/abs/2406.11368

[3] Vishnubhotla, S., et al. (2023). Improving automatic quotation attribution in literary novels. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023). https://aclanthology.org/2023.acl-short.64.pdf

[4] Casanova, E., et al. (2024). XTTS: A massively multilingual zero-shot text-to-speech model. Interspeech 2024. https://arxiv.org/abs/2406.04904

[5] Ruggiero, G., Zovato, E., Di Caro, L., & Pollet, V. (2021). Voice cloning: A multi-speaker text-to-speech synthesis approach based on transfer learning. arXiv preprint arXiv:2102.05630. https://arxiv.org/abs/2102.05630

[6] Casanova, E., et al. (2024). XTTS: Taking TTS to the next level. Coqui Blog. https://coqui.ai/blog/tts/xtts_taking_tts_to_the_next_level

[7] Zhang, Y., et al. (2019). YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion. arXiv preprint arXiv:2112.02418. https://arxiv.org/abs/2112.02418

[8] Neekhara, P., et al. (2021). Expressive neural voice cloning. Proceedings of Machine Learning Research, 157, 1–12. https://proceedings.mlr.press/v157/neekhara21a/neekhara21a.pdf

[9] Reddy, V. K. (2023). Implementation of novel voice cloning method based on deep learning techniques. In Studies in Systems, Decision and Control (Vol. 571, pp. 239–252). Springer. https://link.springer.com/chapter/10.1007/978-3-031-75771-6_20

[10] Coqui.ai. (n.d.). ⓍTTS: TTS 0.22.0 documentation. https://docs.coqui.ai/en/stable/models/xtts.html

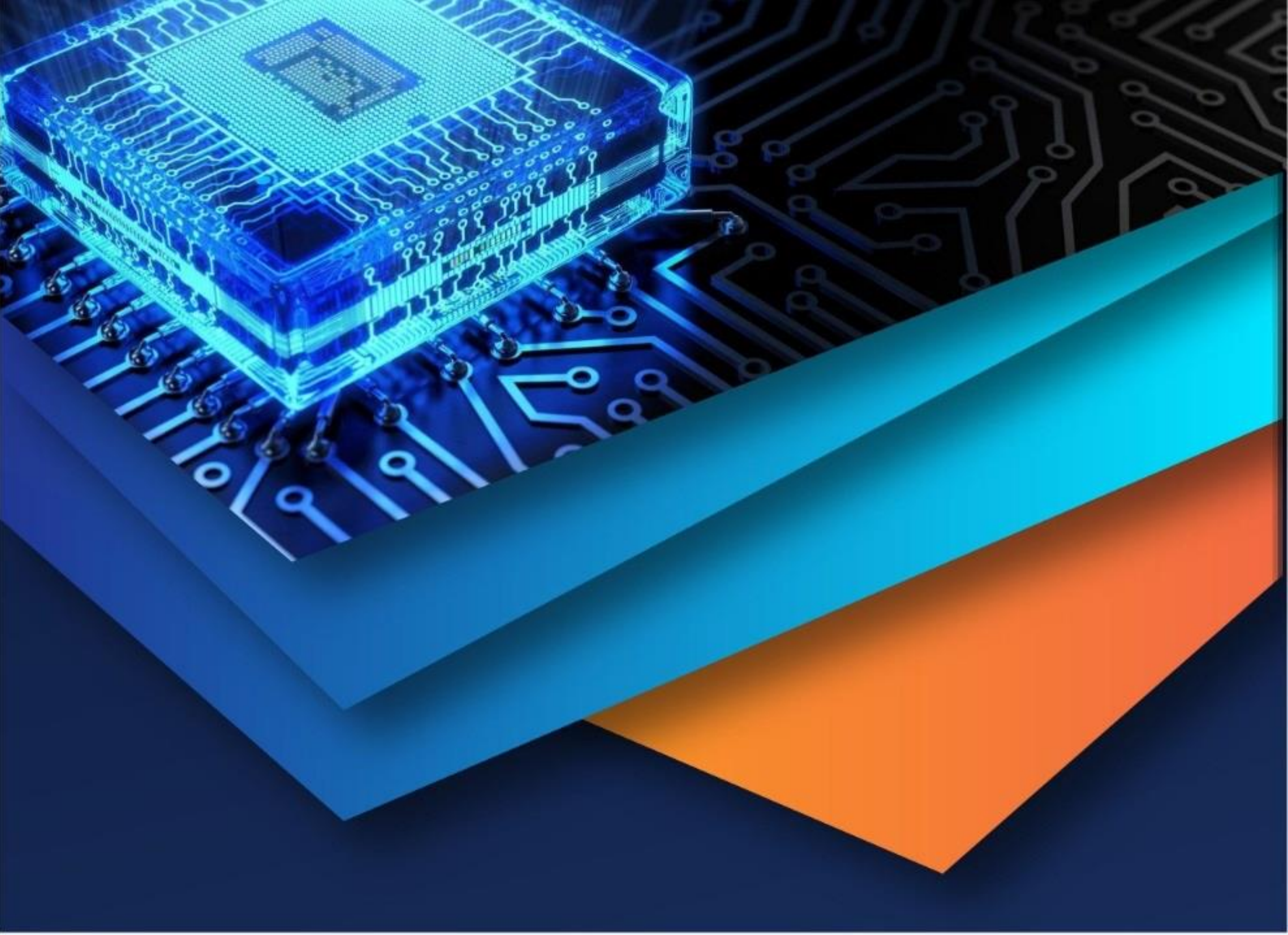
[11] Wang, Y., et al. (2017). Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135. https://arxiv.org/abs/1703.10135

[12] Ren, Y., et al. (2020). FastSpeech 2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558. https://arxiv.org/abs/2006.04558

[13] Zhang, Y., et al. (2021). Boosting multi-speaker expressive speech synthesis with semi-supervised learning. arXiv preprint arXiv:2310.17101. https://arxiv.org/abs/2310.17101

[14] Ramesh, A., et al. (2021). VALL-E: Zero-shot text-to-speech synthesis. arXiv preprint arXiv:2301.02111. https://arxiv.org/abs/2301.02111

[15] Liu, J., et al. (2023). MParrotTTS: Multilingual multi-speaker text to speech synthesis in low-resource settings. arXiv preprint arXiv:2305.11926. https://arxiv.org/abs/2305.11926



doi crossref

10.22214/IJRASET

INDEX COPERNICUS

45.98

ISRA JIF

IMPACT FACTOR: 7.129

SJIFactor

Scientific Journal Impact Factor

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

ISI

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY