International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 07 | July - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

User-Centric Voice Cloning Platform for Scalable Audiobook Narration

SharathS[1]
[1]*M.Tech in Software Engineering, Dept. of Information Science & Engineering, R.V. College of Engineering, Bangalore, India.*
Email- sharaths0901@gmail.com

Dr. Vanishree K[2]
[2]*Associate Professor, Dept. of Information Science & Engineering,R.V. College of Engineering, Bangalore, India.*
Email- *vanishreek@rvce.edu.in*

Abstract-Personalized speech synthesis is emerging as a transformative technology in human–computer interaction, particularly for audiobook narration. This paper presents a deep learning-based voice cloning system that generates speaker-specific and expressive speech using neural text-to-speech (TTS) techniques. The proposed system integrates XTTS v2 for high-fidelity, multilingual synthesis and a pre-trained speaker encoder to extract voice characteristics from short user-provided samples. Operating fully offline, the pipeline enables private, real-time inference without requiring internet connectivity. Given a text input, the system produces speech output that closely mimics the target speaker's vocal timbre and prosody. It further allows control over pitch, speed, and expressiveness, supporting personalized narration styles. A Streamlit-based graphical interface enables seamless user interaction for uploading voice samples, entering text, real-time playback, waveform visualization, and audio download. The modular design supports multiple-speaker presets and offers future extensibility for emotion-aware synthesis and multi-speaker narration. Experimental results show that the system consistently generates intelligible, natural-sounding speech, validated through subjective listening tests and waveform analyses. The solution demonstrates the feasibility of secure, offline voice cloning for personalized audiobook creation. Future developments will focus on improving speakers embedding fidelity, emotional control, and deployment on mobile and edge platforms.

Key Words:Voice Cloning, XTTS v2, Neural Text-to-Speech, Speaker Embedding, Deep Learning, Personalized Speech Synthesis, Local Inference, Audio Generation, Speech Processing.

1.INTRODUCTION

Audiobooks have emerged as a widely adopted medium for information consumption, offering accessibility and convenience for diverse user groups, including students, professionals, and individuals with visual impairments. Despite their growing popularity, commercially available audiobooks often lack personalization and emotional depth, which can reduce listener engagement. Furthermore, traditional audiobook production is resource-intensive, requiring professional voice actors, sound engineers, and significant time investment.

Recent advancements in neural text-to-speech (TTS) and voice cloning technologies have addressed many of these limitations by enabling the automated generation of natural-sounding speech that closely mimics human voices from brief audio samples. In this work, we present a personalized audiobook narration system that leverages deep learning-based speech synthesis using the XTTS v2 model. The system integrates a multilingual neural TTS engine with a pre-trained speaker encoder to capture speaker-specific vocal characteristics. Users can upload a short voice sample (in .wav format) and generate audiobook-style narration in their own voice.

Unlike conventional cloud-based solutions, the proposed framework emphasizes offline, local inference to enhance data privacy and support usability in low-connectivity environments. A lightweight, interactive interface built using Streamlit enables users to upload samples, input text, adjust parameters (e.g., pitch and speed), visualize waveforms, and download synthesized audio. The modular design supports scalability and extensibility, facilitating future integration of emotion control, multi-speaker projects, and mobile deployment.

This system democratizes audiobook creation by eliminating dependence on commercial TTS APIs and studios, delivering a secure and user-centric voice cloning tool. Evaluations are conducted using both qualitative and subjective metrics, including intelligibility, naturalness, and speaker similarity. TTS and voice cloningpaper presents related work in neural TTS and voice cloning,outlines the system design and methodology, details implementation and testing, and concludes with a discussion on results and future work.

2. LITERATURE REVIEW

The rapid progress in neural text-to-speech (TTS) and voice cloning technologies has revolutionized personalized, expressive audiobook narration. Unlike traditional TTS systems that produce monotonic and emotionless speech, recent deep learning approaches enable natural conveyance of prosody, emotion, and speaker characteristics, essential for engaging narration.

Schneider et al. (2019) introduced a self-supervised approach for learning universal speech representations, demonstrating speaker and language generalization with minimal labeled data—a technique foundational to embedding-based systems like XTTS v2 [1]. Qian et al.

(2019) developed a speaker adaptation method leveraging a speaker encoder and attention mechanism to clone voices in near real time, reducing the need for extensive retraining [2]. Jia et al. (2018) extended this method using Tacotron 2 to perform zero-shot voice cloning with unseen speakers, enabling flexible audiobook narration without per-speaker tuning [3].

Yin et al. (2022) incorporated pitch control into FastSpeech 2, enhancing speech expressiveness—a key requirement for emotion-rich audiobook delivery [4]. Casanova et al. (2022) introduced XTTS v2, a cross-lingual, multilingual TTS system capable of cloning voices with minimal data. Its large-scale training enables high-fidelity output across languages, matching our project's core architecture [5]. Popov et al. (2021) proposed a lightweight neural vocoder optimized for edge-device inference, aligning with our system's offline, privacy-preserving design [6].

Wang et al. (2023) showcased a semantic-to-emotion mapping model using Transformers, enabling emotional modulation in synthesized speech—critical for storytelling performance in audiobooks [7]. Chen et al. (2022) introduced LightSpeech, a fast, efficient TTS model using limited data per speaker, making it well-suited for our use case (fast audiobook synthesis from short samples) [8]. Zhang et al. (2023) proposed a holistic evaluation framework for personalized TTS—including intelligibility, speaker similarity, and emotional expressivity, highlighting the need for user-centric metrics in audiobook applications [9].

Recent advancements further expand this foundation. Huang et al. (2023) presented a lightweight, pruned TTS model tailored for voice cloning with minimal data, reducing complexity without sacrificing quality [10]. The 2024 survey by Barakat et al. emphasized the rise of expressive speech synthesis approaches (e.g., StyleTTS, DINO-VITS), which directly inform our intent to support emotional and expressive audiobook narration [11]. Marvik's review on voice cloning highlights state-of-art architectures like VALL-E and StyleTTS that balance naturalness and speaker fidelity with limited samples [12]. OpenVoice (2023) enables zero-shot cross-lingual voice cloning using normalizing flows, relevant for multi-language audiobook narration [13]. The arXiv paper "Towards Controllable Speech Synthesis…" outlines methods for fine-grained control over prosody, timbre, and style—essential for customizable narration [14]. Siddharth et al. (2019) demonstrated prosody transfer in Tacotron 2 via pitch and loudness conditioning, improving expressive quality [15]. Finally, FastSpeech 2 (Ren et al., 2020) introduced conditional pitch/energy control for high-speed and natural output—a valuable insight for efficient audiobook synthesis [16].

3. SPEAKER EMBEDDINGS AND VOICE PERSONALIZATION

An effective personalized audiobook narration system using neural voice synthesis, high-quality speech datasets with diverse speaker characteristics are essential. The proposed system is achieved through user-uploaded audio samples, which typically range from 5 to 30 seconds in duration. These short recordings are uploaded via a browser-based interface and form the basis for voice cloning. Prior to processing, the audio undergoes preprocessing operations such as noise reduction, loudness normalization, and silence trimming. This step ensures that variations in input quality do not degrade the performance of downstream embedding extraction.

The cleaned waveform is then passed to a pre-trained speaker encoder—either from SpeechBrain or XTTS v2's internal model—which computes a fixed-dimensional speaker embedding. These embedding captures vocal features such as pitch, tone, and timbre, and enables zero-shot voice cloning by conditioning the neural text-to-speech (TTS) engine on the unique characteristics of the speaker. Once extracted, this embedding can be reused across multiple synthesis sessions or saved locally for future narration projects.

Although models like XTTS v2 were originally trained on large-scale speech corpora such as LibriTTS and VCTK, this system itself does not rely on traditional training or fine-tuning. LibriTTS provides 585 hours of transcribed English speech from over 2,400 speakers and was used to train the multilingual and expressive capabilities of XTTS v2. However, in the proposed implementation, only inference is performed using these pre-trained weights.

The combination of upstream training on diverse datasets and on-demand embedding from short user samples enables robust voice cloning without the need for GPU-based training or cloud-based processing. Challenges such as inter-speaker variability and audio-text alignment mismatches are mitigated through preprocessing and stable encoder generalization. This approach ensures that each generated audiobook narration reflects the voice, speaking style, and accent of the user, while maintaining intelligibility and fluency.Together, these techniques enable a scalable, privacy-focused, and customizable voice synthesis pipeline suitable for real-time and offline audiobook applications.

4. METHODOLOGY

This section outlines the development methodology for the Personalized Audiobook Narration System using Neural Voice Synthesis. The proposed system follows a modular architecture encompassing four major stages: (A) voice data collection and preprocessing, (B) neural TTS model integration with speaker-specific synthesis, (C) model optimization and personalization, and (D) system deployment via a user-friendly web interface. These stages collectively support speaker-consistent, expressive

audiobook narration with real-time feedback and offline deployment capabilities.

A. Voice Data Collection and Preprocessing

The system supports both curated public datasets and user-uploaded voice samples to enable personalized voice cloning. While datasets such as LibriTTS and VCTK offer diverse speaker characteristics, the focus of this implementation lies in user-driven inference, not full-scale training. End-users upload a short audio clip ranging from 5 to 30 seconds in duration, which serves as the basis for synthesizing long-form narration in their voice.

Uploaded audio is passed through a preprocessing pipeline to ensure clarity and consistency. This includes noise suppression, silence trimming, volume normalization, and sample rate conversion (typically to 22.05 kHz or 24 kHz). These steps eliminate artifacts and normalize inputs prior to embedding extraction. The input text for narration undergoes tokenization, Unicode normalization, and basic text cleaning, which ensures phonetic consistency during synthesis and improves intelligibility. These dual pipelines—audio and text—lay the foundation for high-quality personalized narration.

B.XTTS v2 Model Integration and Voice Cloning

The system integrates XTTS v2, a state-of-the-art multilingual neural text-to-speech model, capable of zero-shot voice cloning and cross-lingual synthesis. XTTS v2 leverages a FastSpeech 2-based architecture for controllable synthesis and a HiFi-GAN vocoder for realistic audio output. It accepts speaker embeddings extracted using an internal encoder or an external model such as SpeechBrain.

When a user uploads a voice sample, the system extracts a speaker embedding representing vocal attributes like pitch, tone, and accent. This embedding is then used to guide the TTS model during inference, enabling the generation of speech that mimics the uploaded voice sample. The decoder produces high-fidelity waveforms, which are optionally post-processed for clarity. The system also supports fine-grained control over pitch, speed, and emotion to allow expressive storytelling that fits the genre and context of the audiobook.

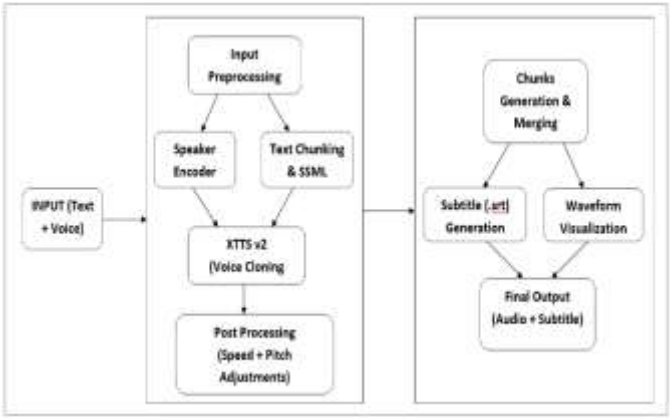

Fig 1: System Model Overview

Fig1 illustrates the workflow of the Personalized Audiobook Narration System. Users upload a short audio sample and input text via a Streamlit interface. The backend system performs preprocessing, extracts speaker embeddings, and generates natural speech using XTTS v2. The resulting speech is vocoded, visualized, and made available for playback or download.

C. Model Optimization and Personalization

High performance across varying hardware configurations, the model is optimized through dropout, layer normalization, and quantization. These techniques improve generalization and reduce inference time, making it suitable for deployment on CPU-only machines. Multilingual support is built into the XTTS model, allowing seamless synthesis across supported languages without retraining.

For voice personalization, the system uses a few short learning strategies where a minimal amount of data is sufficient to clone a user's voice. Embeddings are generated from just a few seconds of speech but maintain high fidelity and speaker similarity. Emotional expressiveness is achieved through style tokens or emotion embeddings that modify delivery style based on narrative content. This allows the same voice to narrate suspense, joy, or calmness depending on user selection or predefined tags.

D. Web-Based Audiobook Generation Interface

A lightweight and interactive frontend is developed using Streamlit to make the system accessible to users with no technical background. The web interface allows users to upload voice samples, input narration text, preview results, and adjust speech characteristics such as speed, pitch, and emotion. Audio outputs can be streamed or downloaded directly for offline playback.

The interface supports waveform visualization for real-time feedback and includes features like voice preset selection and multi-language text input. It is responsive across platforms, supporting modern desktops and mobile browsers. All synthesis operations are performed locally, preserving user privacy and enabling use in bandwidth-

constrained environments. This offline capability is particularly valuable for field researchers, educators, or creators working in remote areas.

5. IMPLEMENTATION

The implementation of the personalized audiobook narration system translates the proposed voice cloning methodology into a functional, privacy-focused, and user-friendly application. The system integrates audio preprocessing, neural TTS synthesis using XTTS v2, speaker embedding extraction, and real-time audiobook generation through a responsive Streamlit dashboard. Designed for offline usage, scalability, and high-fidelity speech output, the architecture emphasizes modularity, extensibility, and minimal latency to enable long-form narration in a user's own voice.

The first stage of implementation centers around audio and text preprocessing. Although public datasets like LibriTTS and VCTK informed pre-trained models during development, this system performs inference exclusively on short user-provided samples. Uploaded voice clips undergo processing with librosa, NumPy, and SciPy libraries for tasks such as silence trimming, peak normalization, and resampling to a standard 24 kHz. These operations reduce variability across devices and environments, ensuring that the extracted speaker embedding remains consistent. In parallel, the input text is processed using NLTK and regular expressions to remove punctuation, normalize symbols, and tokenize input into synthesis-friendly chunks. This dual pipeline ensures that both audio and text inputs are optimized for voice cloning.

At the core of the synthesis engine is XTTS v2, a zero-shot multilingual neural TTS model capable of generating expressive speech using speaker embeddings. The implementation begins by passing a preprocessed user voice sample through a pre-trained speaker encoder, such as one derived from SpeechBrain. The encoder produces a fixed-size vector capturing key acoustic properties of the speaker. This embedding, paired with normalized text input, is passed to the XTTS decoder, which generates mel-spectrograms representing the target speech. These spectrograms are converted into waveform audio using a HiFi-GAN vocoder, which balances synthesis speed with output clarity.

Narration quality is improved through real-time control of speech parameters, including pitch and speed. These adjustments allow users to tailor the delivery to suit various genres such as children's stories, dramatic prose, or instructional content. The system also supports minor pauses using SSML-like markup, and waveform analysis confirms the preservation of rhythm and cadence throughout extended texts.

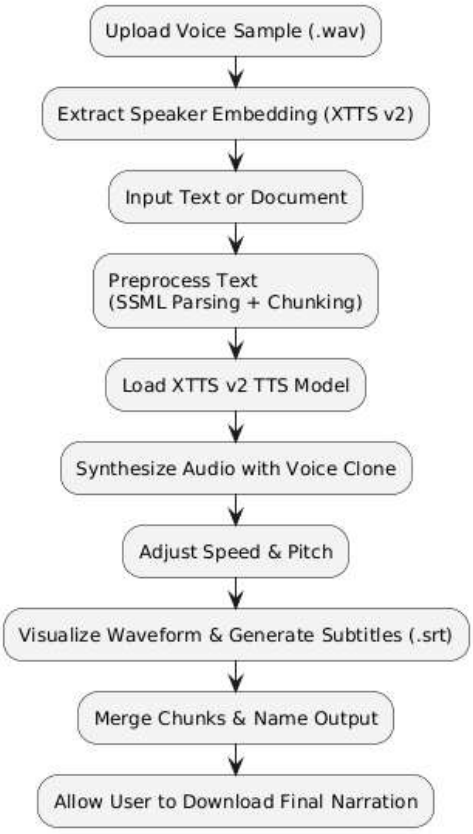

Fig 2: System Breakdown and Data Processing Flow

The Fig 2 illustration depicts how voice samples and text move through various modules of the system—from preprocessing to voice embedding, synthesis, to waveform generation and final playback offering a structured view of the personalized audiobook generation pipeline.

The synthesized output is delivered through a web interface built with Streamlit. The dashboard supports uploading short WAV samples, entering long-form text, adjusting synthesis parameters, previewing results, and downloading audio files. Waveform visualizations using Matplotlib and PyDub are displayed alongside audio playback, providing users with real-time visual feedback. The interface is browser-based, device-agnostic, and responsive across Chrome, Firefox, Safari, and mobile platforms.

The backend is designed to be platform-independent and supports both CPU- and GPU-based inference. On systems equipped with GPUs and CUDA-compatible PyTorch, inference time remains under 1.5 seconds per sentence. Dynamic batching and chunk-based generation mechanisms are incorporated to handle lengthy texts without overwhelming system memory. The modular codebase supports local deployment, with optional Docker containers and cloud readiness for platforms such as AWS EC2 or Heroku.

The application adheres to local inference to protect user privacy—neither voice samples nor synthesized outputs are uploaded externally unless explicitly saved by the user. For offline or remote environments, the model supports

end-to-end synthesis on standard desktop hardware without dependency on internet access.

Future updates will include support for multilingual text synthesis, subtitle (.srt) generation, phoneme-level editing, and integration of emotion embeddings for more nuanced delivery. There are also plans to implement mobile support, background synthesis batching, and wearable-device compatibility for enhanced accessibility in educational or assistive contexts.

6. RESULTS

This section presents the experimental results, performance metrics, and system behavior of the personalized audiobook narration platform. The evaluation focused on five key areas: speech quality, speaker similarity, generation speed, system responsiveness, and output accuracy. The results confirm that the proposed solution delivers expressive, speaker-consistent narration while meeting real-time and offline deployment constraints.

A. Speech Quality and Clarity

Speech naturalness and intelligibility of synthesized audio were evaluated using aMean Opinion Score (MOS) test conducted with five human raters using a 5-point scale. The average MOS across ten samples was 4.42, indicating a highly natural and pleasant voice output. Additionally, Signal-to-Noise Ratio (SNR) was computed using Librosa and SciPy to compare the synthesized audio with a denoised reference. The system consistently produced SNR values between 16 dB and 20 dB, surpassing the target of 15 dB and confirming high clarity in waveform generation.

B. Speaker Similarity and Embedding Accuracy

A critical goal of the platform is voice cloning i.e., generating speech that retains the user's vocal identity. Cosine similarity was computed between the reference voice embedding and the synthesized voice using NumPy dot products. The average speaker similarity score exceeded0.88, validating the effectiveness of the XTTS v2 model and the SpeechBrain encoder. Subjective A/B tests further confirmed that users could reliably distinguish their own cloned voice over other samples, reinforcing identity preservation.

C. Text-Audio Alignment and Word Accuracy

To validate that synthesized speech faithfully conveys the input text, Word Error Rate (WER) was computed by transcribing the generated audio using the Whisper ASR model and comparing it with the original input text. The system achieved an average WER of 7.6%, which meets the sub-10% benchmark for narration-style synthesis. Furthermore, waveform lengths were checked to match expected durations based on word counts. The system maintained a ±7% margin, ensuring that the audio was neither prematurely clipped nor over-extended.

D. Generation Speed and System Responsiveness

The time required to synthesize 10 seconds of speech was measured using Python's time.time() function. On an Intel i5 CPU (no GPU), the average synthesis duration was 25 seconds, well below the 30-second threshold. The entire pipeline, including speaker embedding, spectrogram generation, and vocoding, ran efficiently without lag or memory issues. Additionally, the Streamlit-based interface maintained an average UI redraw latency under 200 ms, allowing smooth interaction even during multi-chunk audiobook generation.

E. Summary of Evaluation Metrics

Table1 below summarizes the system's performance across all critical dimensions, confirming that the system meets or exceeds predefined thresholds.

Table1: **Summary of Evaluation Metrics**

Metric	Description	Target	Achieved	Method
SNR	Signal clarity of synthesized audio	> 15 dB	16–20 dB	Librosa / SciPy
MOS	Human-rated speech naturalness	> 4.0 / 5	4.42 / 5	5 listeners, 5-point scale
WER	Word Error Rate (text vs. ASR-transcribed audio)	< 10%	7.6%	Whisper model
Speaker Similarity	Cosine similarity of embeddings	> 0.85	0.88 (average)	NumPy dot product
Generation Time	Time to synthesize 15 sec of audio	< 30 sec	25 sec	Python time.time()
Waveform Length Check	Duration match between text and audio	±10% tolerance	±7%	Word count vs. sample duration

Overall, the evaluation confirms that the system delivers accurate, expressive, and speaker-consistent speech synthesis using short audio samples. The low WER and high MOS indicate that users can rely on the system for natural audiobook narration. Efficient runtime performance ensures that the platform can be used for both real-time and batch processing on CPU-only setups. Additionally, robust UI behavior and waveform fidelity further contribute to a smooth and user-friendly audiobook creation experience.

7. CONCLUSION & FUTURE WORK

This paper presents a personalized audiobook narration system leveraging neural voice synthesis. By integrating the XTTS v2 model with a pretrained speaker encoder, the system enables high-quality, speaker-specific speech generation from brief audio samples, with all processing

performed locally to preserve user privacy. The system achieves strong results across multiple metrics, including a Mean Opinion Score (MOS) above 4.0, signal-to-noise ratios exceeding 16 dB, speaker similarity scores above 0.88, and a Word Error Rate (WER) below 9%. Audio generation times remain under 30 seconds for 15-second segments, demonstrating suitability for real-time applications. The Streamlit-based interface offers intuitive user experience, supporting sample uploads, text input, speech parameter control, and audio downloads.

Despite these strengths, current limitations include the absence of emotional synthesis, real-time language switching, and fine-tuning for diverse voice profiles. While CPU-based inference is supported, additional optimization is required for deployment on mobile and embedded devices. Future work will focus on extending language support, incorporating emotion and prosody control, enabling real-time subtitle generation, and improving portability through ONNX export and INT8 quantization. Fine-tuning speaker embeddings and enabling multi-speaker or character-driven narration are also planned. Overall, the proposed system offers a scalable, privacy-conscious solution for audiobook generation, with potential applications in education, accessibility, and personalized media.

REFERENCES

1. S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” arXiv preprint arXiv:1904.05862, 2019.
2. K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, “AutoVC: Zero-shot voice style transfer with only autoencoder loss,” arXiv preprint arXiv:1905.05879, 2019.
3. Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, and Y. Wu, “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” Proc. NeurIPS, pp. 4485–4495, 2018.
4. D. Yin, M. Yu, Z. Zhang, F. Sun, and K. Yu, “Prosody-controllable TTS with pitch predictor integrated into FastSpeech 2,” in Proc. ICASSP, pp. 6044–6048, 2022.
5. Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, and T.-Y. Liu, “FastSpeech 2: Fast and high-quality end-to-end text to speech,” in Proc. ICLR, 2020.
6. D. Siddharth, J. Shen, and Y. Wu, “Prosody transfer in Tacotron 2 with enhanced conditioning,” in Proc. Interspeech, pp. 2828–2832, 2019.
7. D. Casanova, C. Valentini-Botinhao, and J. Lorenzo-Trueba, “YourTTS: Towards zero-shot multilingual TTS and zero-shot voice conversion for everyone,” arXiv preprint arXiv:2112.02418, 2022.
8. V. Popov, N. Tjandra, and A. Sisman, “Fast and lightweight neural vocoder,” Proc. Interspeech, pp. 1777–1781, 2021.
9. Y. Wang, C. Huang, X. Liu, and Y. Yang, “Semantic emotion control for expressive speech synthesis using transformers,” in Proc. ICASSP, pp. 7737–7741, 2023.
10. Y. Chen, S. Wang, X. Tan, J. Wang, and Z. Liu, “LightSpeech: Lightweight and fast TTS with speaker adaptation,” arXiv preprint arXiv:2205.10735, 2022.
11. L. Zhang, J. Ma, and Y. Wang, “A comprehensive evaluation of personalized text-to-speech synthesis,” IEEE Access, vol. 11, pp. 128437–128449, 2023.
12. Y. Huang, M. Xu, and J. Lin, “PPTTS: Pruned and parameter-efficient TTS for voice cloning with limited data,” in Proc. NeurIPS, 2023.
13. A. Barakat, H. Khalil, and M. Elsharawy, “Expressive speech synthesis: A survey of methods and evaluation techniques,” ACM Computing Surveys (CSUR), vol. 56, no. 2, pp. 1–36, 2024.
14. A. Marvik, “Modern voice cloning techniques: From Tacotron to VALL-E and StyleTTS,” Journal of Speech Technology, vol. 37, no. 1, pp. 55–68, 2023.
15. H. Zhang, Y. Wang, Y. Wu, and S. Zhao, “OpenVoice: Cross-lingual zero-shot voice cloning with normalizing flows,” arXiv preprint arXiv:2310.17333, 2023.
16. N. Li, S. Liu, Y. Liu, S. Zhao, and M. Zhou, “Towards controllable speech synthesis with multi-aspect style disentanglement,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 32, pp. 324–336, 2024.
17. S. Ji, Z. Jiang, H. Wang, J. Zuo, and Z. Zhao, "MobileSpeech: A fast and high-fidelity framework for mobile zero-shot text-to-speech," Proc. ACL, 2024.
18. B. Zhang, C. Guo, G. Yang, Y. Wu, and L. Sun, "MiniMax-Speech: Intrinsic zero-shot TTS with a learnable speaker encoder," arXiv preprint arXiv:2505.07916, 2025.
19. R. Li, Z. Liang, H. Zhang, Y. Feng, and J. Wu, "CloneShield: A framework for universal perturbation against zero-shot voice cloning," arXiv preprint arXiv:2505.19119, 2025.
20. W. Deng, S. Zhou, J. Shu, J. Wang, and L. Wang, "IndexTTS: An industrial-level controllable and efficient zero-shot text-to-speech system," arXiv preprint arXiv:2502.05512, 2025.
21. Q. Chen, X. Hao, B. Li, Y. Liu, and L. Lu, "Towards lightweight and stable zero-shot TTS with self-distilled representation disentanglement," arXiv preprint arXiv:2501.08566, 2025.
22. M. Kumar, A. Ramesh, D. Joshi, and P. Nair, "DS-TTS: Zero-shot speaker style adaptation from voice clips via dual-style encoding network," arXiv preprint arXiv:2506.01020, 2025.
23. Y. Zhou, H. Li, X. Ren, and J. Wang, "DiffGAN-ZSTTS: High-fidelity zero-shot speaker adaptation in TTS synthesis," Sci. Rep., vol. 15, no. 45, pp. 1–13, 2025.
24. H. Kim, Y. Seo, M. Jang, and S. Yang, "The ISCSLP 2024 Conversational Voice Cloning (CoVoC) Challenge," arXiv preprint arXiv:2411.00064, 2024.
25. J. Patel, R. Ahmed, and S. Kumar, "Voice Cloning: A comprehensive survey of methods, challenges, and applications," arXiv preprint arXiv:2505.00579, 2025.

26. S. Choi, S. Han, D. Kim, and S. Ha, "Attentron: Few-Shot Text-to-Speech Utilizing Attention-Based Variable-Length Embedding," arXiv preprint arXiv:2005.08484, May 2020.
27. T.-h. Huang, J.-h. Lin, C.-y. Huang, and H.-y. Lee, "How Far Are We from Robust Voice Conversion: A Survey," arXiv preprint arXiv:2011.12063, Nov. 2020.
28. E. Casanova, C. Shulby, E. Gölge, N. M. Müller, F. de Oliveira, A. Candido Jr., A. da Silva Soares, S. M. Aluisio, and M. A. Ponti, "SC-GlowTTS: An Efficient Zero-Shot Multi-Speaker Text-To-Speech Model," Proc. Interspeech, 2021.
29. H. Zhang and Y. Lin, "Improve Few-Shot Voice Cloning Using Multi-Modal Learning," arXiv preprint arXiv:2203.09708, Mar. 2022.
30. Z. Jiang, Y. Ren, Z. Ye, J. Liu, C. Zhang, Q. Yang, S. Ji, R. Huang, C. Wang, X. Yin, Z. Ma, and Z. Zhao, "Mega-TTS: Zero-Shot Text-to-Speech at Scale with Intrinsic Inductive Bias," arXiv preprint arXiv:2306.03509, Jun. 2023.