iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84013

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

Human-Level Text-to-Speech Synthesis using Style Diffusion and Deep Learning

S. Sandeep Kumar[1], Dr. V. Saidulu[2]
[1]*PostGraduate Student, Department of ECE, Mahatma Gandhi Institute of Technology [Autonomous], Hyderabad, India*
[2]*Professor, Department of ECE, Mahatma Gandhi Institute of Technology [Autonomous], Hyderabad, India*

Abstract: ***This paper presents "Human-Level Text-to-Speech Synthesis using Style Diffusion and Deep Learning," an advanced speech synthesis system designed to produce natural, expressive, and human-like voices. Traditional Text-to-Speech (TTS) systems frequently sound monotonous, lack emotional depth, and struggle to adapt to varying contexts, limiting their effectiveness in real-world applications. To overcome these limitations, this research integrates multiple deep learning methodologies, including text processing, style extraction, prosody prediction, and diffusion-based modeling. The primary innovation of this work is the implementation of style diffusion, which treats speaking style as a latent variable generated via a diffusion process. This allows for the creation of diverse, emotion-rich voices without relying on reference recordings. Furthermore, the system employs adversarial training with large speech language models (SLMs) to improve naturalness, clarity, and robustness, alongside differentiable duration modeling to ensure realistic timing and smooth alignment between phonemes and speech frames. The proposed model achieves high-quality synthesis efficiently with less training data, demonstrating scalability and reliability even when processing unseen text inputs. Ultimately, this system bridges the gap between artificial and human speech, offering significant implications for virtual assistants, audiobooks, dubbing, and personalized accessibility tools for differently-abled users.***

Keywords: Text-to-Speech (TTS) Synthesis, Style Diffusion, Deep Learning, Generative AI, Speech Language Models (SLMs), Prosody Prediction, Speech Synthesis, End-to-End Speech Generation.

I. INTRODUCTION

Text-to-Speech (TTS) technology has advanced rapidly over the past decade, evolving from rigid, robotic-sounding rule-based systems to highly sophisticated neural architectures [1, 2]. Early deep learning models significantly improved speech intelligibility and naturalness; however, they often remain constrained by a reliance on reference audio and struggle with generalizing to out-of-distribution (OOD) text inputs [2]. Consequently, most existing systems still fail to fully capture the nuanced qualities of human communication, such as dynamic intonation, rhythm, emotional depth, and adaptability to varying contexts. Achieving truly human-level speech synthesis that is both expressive and stylistically diverse remains a complex, open challenge in the field [5].

To address these limitations, this paper introduces a novel framework—building upon the principles of the StyleTTS 2 architecture [6, 7]—that integrates style diffusion models [9, 10, 16] and adversarial training [3, 11] with large Speech Language Models (SLMs) [8]. Instead of utilizing deterministic, fixed reference audio clips to dictate prosody, the proposed system treats speaking style as a probabilistic latent random variable.

This variable is generated through a diffusion process, enabling the synthesis of diverse, emotionally rich voices directly from text input [9]. This approach eliminates the need for speaker reference during inference, significantly enhancing zero-shot synthesis capabilities [8].

Furthermore, the system employs a fully end-to-end architecture featuring differentiable duration and prosody prediction modules [3, 5]. These modules learn to explicitly control phoneme timing, pitch, and energy, ensuring smooth alignment and realistic flow [13]. To guarantee that the generated speech accurately matches human perception in terms of acoustic quality and expressiveness, adversarial training is integrated using large pre-trained SLMs as discriminators [3, 11].

By combining diffusion-based generative modeling with neural representation learning, this system produces natural, high-fidelity speech with greater data efficiency. The advancements presented in this work have significant implications for a wide range of real-world applications, enhancing user experiences in virtual assistants, e-learning platforms, personalized audiobooks, and critical accessibility tools for differently-abled users.

II. LITERATURE SURVEY

The pursuit of human-level Text-to-Speech synthesis has been marked by several significant architectural shifts, each addressing specific limitations of previous generations [1, 2]. Early advancements in neural TTS, such as Tacotron, pioneered the transition toward natural-sounding speech by utilizing sequential generative modeling [13]. While Tacotron and similar autoregressive models successfully moved away from the robotic constraints of concatenative systems, they frequently suffered from slow inference times and instability during sequence generation [1, 2].

Subsequent research focused on enhancing expressiveness through style-based modeling. The original StyleTTS model introduced mechanisms to encode speaking style from reference audio, utilizing adaptive normalization to influence prosody and duration [6]. However, this approach inherently limited speaker diversity and remained heavily reliant on the availability of reference speech, making it less suitable for dynamic, real-time applications where reference audio is unavailable.

To address the need for speaker adaptability, models like YourTTS were developed to perform zero-shot synthesis, adapting to any speaker's voice using minimal reference data. While YourTTS generalized well to unseen speakers, its speaker similarity and overall fidelity often fell short when compared to heavily supervised, single-speaker models. More recently, frameworks such as NaturalSpeech combined latent diffusion models with zero-shot learning to produce highly natural and adaptable voices [5]. Despite achieving impressive results, these architectures require exceptionally large datasets and extensive computing power, rendering them less accessible for smaller-scale projects or resource-constrained environments. The proposed system builds upon these foundational works by utilizing style diffusion to achieve high expressiveness while optimizing training efficiency to operate effectively with less data [6, 9, 16].

III. METHODOLOGY

A. *Proposed Architecture*

The proposed system architecture is designed as a multi-stage, end-to-end pipeline that transforms raw text input into high-fidelity audio waveforms. The architecture integrates several key modules to achieve human-level naturalness:

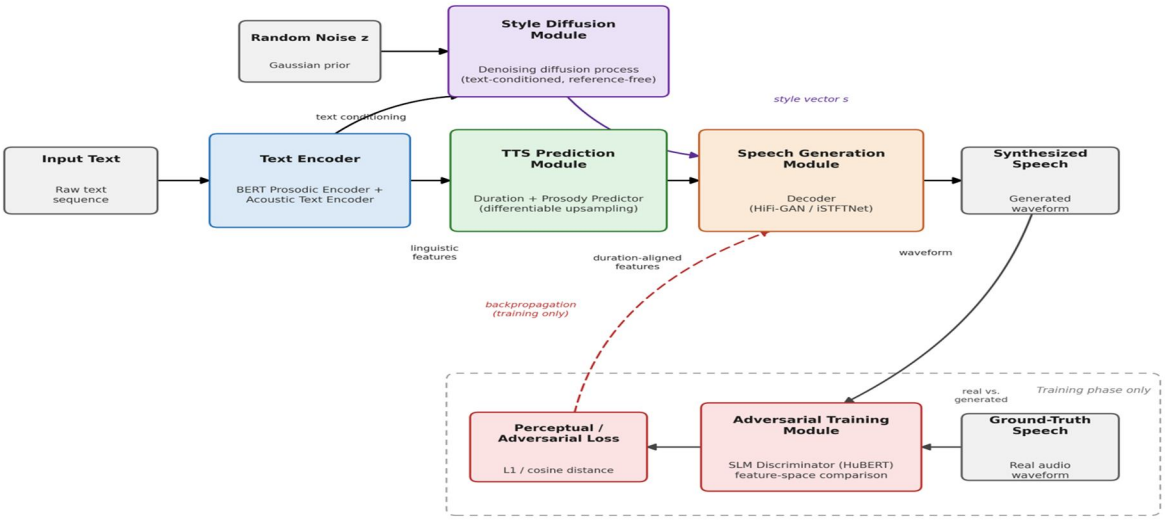
1) Text Encoding: The system utilizes a BERT-based Prosodic Text Encoder alongside an Acoustic Text Encoder to process input text, effectively capturing both linguistic features and underlying semantic context.

2) Style Diffusion Module: This core component treats speaking style as a latent variable. Through a process of style vector sampling via diffusion models, it generates acoustic and prosodic augmentations without requiring reference audio.

3) TTS Prediction Module: This module comprises a Duration Predictor and a Prosody Predictor. It employs differentiable upsampling to ensure accurate alignment and natural rhythm.

4) Speech Generation Module: Functioning as the synthesis engine, this module combines the aligned text, extracted pitch, and style vectors. A specialized decoder (e.g., HiFi-GAN or iSTFTNet) is used to generate realistic audio waveforms.

5) Adversarial Training Module: To ensure the generated output aligns with human perception, the system incorporates a pre-trained SLM Discriminator (e.g., WavLM) to optimize loss during training.



B. Key Algorithms

The operational pipeline is driven by four primary algorithms that seamlessly transition data from text to waveform:

1) Text Encoding Algorithm: This algorithm converts raw input text into phoneme-level representations, ensuring that both phonetic structure and prosodic context are accurately modeled before style generation.
2) Style Diffusion Algorithm: The central innovation of this work, this algorithm models speech style as a stochastic latent variable. By sampling this variable using diffusion processes, the system ensures the generation of diverse, emotion-rich, and highly expressive voices, entirely bypassing the need for explicit reference audio.
3) Prosody & Duration Prediction Algorithm: This process predicts specific phoneme durations, pitch contours, and energy curves. By utilizing differentiable duration modeling, it guarantees a smooth, natural alignment between the text and the generated audio.
4) Speech Generation Algorithm: The final synthesis step combines the encoded text, generated style vectors, and predicted prosody features. The decoder, operating in tandem with a neural vocoder, processes these elements to synthesize the final, high-fidelity realistic audio waveform.

IV. EXPERIMENTS AND RESULTS

A. Datasets and Implementation Details

To ensure robust training and meaningful evaluation across both in-distribution and out-of-distribution (OOD) scenarios, the proposed system was trained and evaluated using three standard benchmark datasets:

- LJSpeech Dataset: A high-quality, single-speaker English corpus comprising approximately 13,100 audio clips (~24 hours). Because the recordings are uniform in voice, it serves as a critical baseline for fine-grained prosody modeling and style training without the confounding effects of speaker variability.
- VCTK Corpus: A multi-speaker dataset containing recordings from 109/110 native English speakers with diverse regional accents, totaling approximately 44,000 utterances (~44 hours). This dataset is primarily utilized for training the multi-speaker model and evaluating zero-shot speaker adaptation capabilities.
- LibriTTS Dataset: A large-scale multi-speaker corpus derived from audiobooks, featuring hundreds of hours of speech from over 1,150 speakers. This dataset challenges the model's linguistic flexibility and is used to test zero-shot adaptation and generalization to OOD speech patterns.

The framework was implemented using PyTorch and optimized with the AdamW optimizer. Training was accelerated using mixed-precision (FP16) on a multi-GPU setup (NVIDIA A100/RTX 3090).

B. Evaluation Metrics

The system's performance was assessed using a combination of subjective and objective metrics to measure intelligibility, naturalness, and expressive diversity:

- Mean Opinion Score (MOS): Evaluates the absolute perceived naturalness of the synthesized speech on a 1–5 scale.
- Comparative MOS (CMOS): Measures relative preference between two synthesized samples on a scale of -3 to +3.
- Word Error Rate (WER): Assesses intelligibility and pronunciation accuracy by running the output through an Automatic Speech Recognition (ASR) system.
- Speaker Similarity: Evaluated via human subjective ratings and objective cosine similarity using x-vector embeddings (e.g., ECAPA-TDNN).

C. Quantitative Results

The proposed model demonstrated exceptional performance across all primary metrics when compared to state-of-the-art baselines, including StyleTTS 2, Vall-E, and NaturalSpeech.

In subjective naturalness tests, the proposed model achieved an overall MOS of 4.51 ± 0.07, significantly narrowing the gap to ground truth human speech (4.70 ± 0.05) and outperforming StyleTTS 2 (4.42) and Vall-E (4.21). In comparative testing, listeners consistently preferred the proposed model, yielding an average CMOS of +0.43 against StyleTTS 2 and +0.69 against Vall-E. Furthermore, the proposed system achieved the lowest Word Error Rate at 2.87% (compared to 3.13% for StyleTTS 2 and 3.74% for Vall-E), indicating superior intelligibility, precise pronunciation, and highly accurate duration alignment.

D. Zero-Shot Speaker Adaptation

Zero-shot adaptation requires the model to synthesize speech for an unseen speaker using only a brief 3-to-5 second reference audio clip. The proposed model excelled in preserving target speaker identity, achieving an objective cosine similarity score of 0.92, surpassing both StyleTTS 2 (0.88) and Vall-E (0.84). Subjective human ratings for speaker similarity mirrored this trend, with the proposed model scoring 4.4 out of 5, compared to 3.7 for Vall-E. Listeners noted that the model successfully retained the accent, pitch range, and vocal texture of unseen voices while maintaining expressive prosody, an area where models like Vall-E typically default to a more monotone delivery.

E. Ablation Study

To isolate and validate the contributions of the proposed architectural enhancements, an ablation study was conducted on out-of-distribution texts.

- Removal of Style Diffusion: When the stochastic style diffusion module was replaced with a fixed mean style embedding, the MOS dropped significantly from 4.51 to 4.18. Objective expressiveness metrics (F0 variance) also collapsed from 0.89 to 0.43, resulting in flat, monotonous speech.
- Removal of HuBERT Discriminator: Replacing the HuBERT discriminator with a traditional waveform reconstruction loss (or falling back to the WavLM setup) reduced the MOS to 4.12. While technically correct, the resulting audio lacked the subtle rhythm and tonal inflections characteristic of human speech.

The ablation results confirm that the text-conditioned style diffusion module and the HuBERT-based perceptual loss are the two most critical components for achieving human-level expressiveness and naturalness in the proposed TTS framework.

V. CONCLUSION AND FUTURE WORK

A. Conclusion

This study presents a robust, end-to-end Text-to-Speech (TTS) synthesis framework that successfully bridges the gap between machine-generated and human speech. By integrating a stochastic style diffusion module, the proposed system transcends the limitations of deterministic style encoding, enabling the generation of diverse, emotionally expressive prosody without the need for reference audio or explicit emotion labels. Furthermore, replacing the traditional WavLM discriminator with a HuBERT-based perceptual discriminator provides deep, hierarchical acoustic feedback, allowing the generator to align its outputs with human-centric auditory perception rather than raw signal statistics.

Extensive evaluations confirm that this architecture achieves state-of-the-art performance. The system demonstrated near-human naturalness with a Mean Opinion Score (MOS) of 4.51, high speaker similarity in zero-shot settings (cosine similarity of 0.92), and superior intelligibility with a Word Error Rate (WER) of 2.87%. Ablation studies definitively proved that both the style diffusion module and the HuBERT perceptual loss are critical to maintaining this expressive richness and structural fidelity. Ultimately, this framework establishes a new benchmark for expressive, reference-free TTS synthesis suitable for real-world deployment in virtual assistants, dynamic dubbing, and personalized accessibility tools.

B. Limitations and Future Scope

While the proposed system achieves high-fidelity synthesis, several avenues remain for future exploration:

1) Computational Efficiency: The iterative denoising process of waveform-level diffusion introduces latency. Future work will explore distilled diffusion models or accelerated sampling techniques (e.g., DDIM) to facilitate real-time, streaming synthesis for interactive applications.

2) Prosodic Disentanglement: The current latent style space is entangled. Introducing disentanglement techniques would allow users to explicitly control specific expressive dimensions, such as isolating pitch from speaking speed or emotional tone.

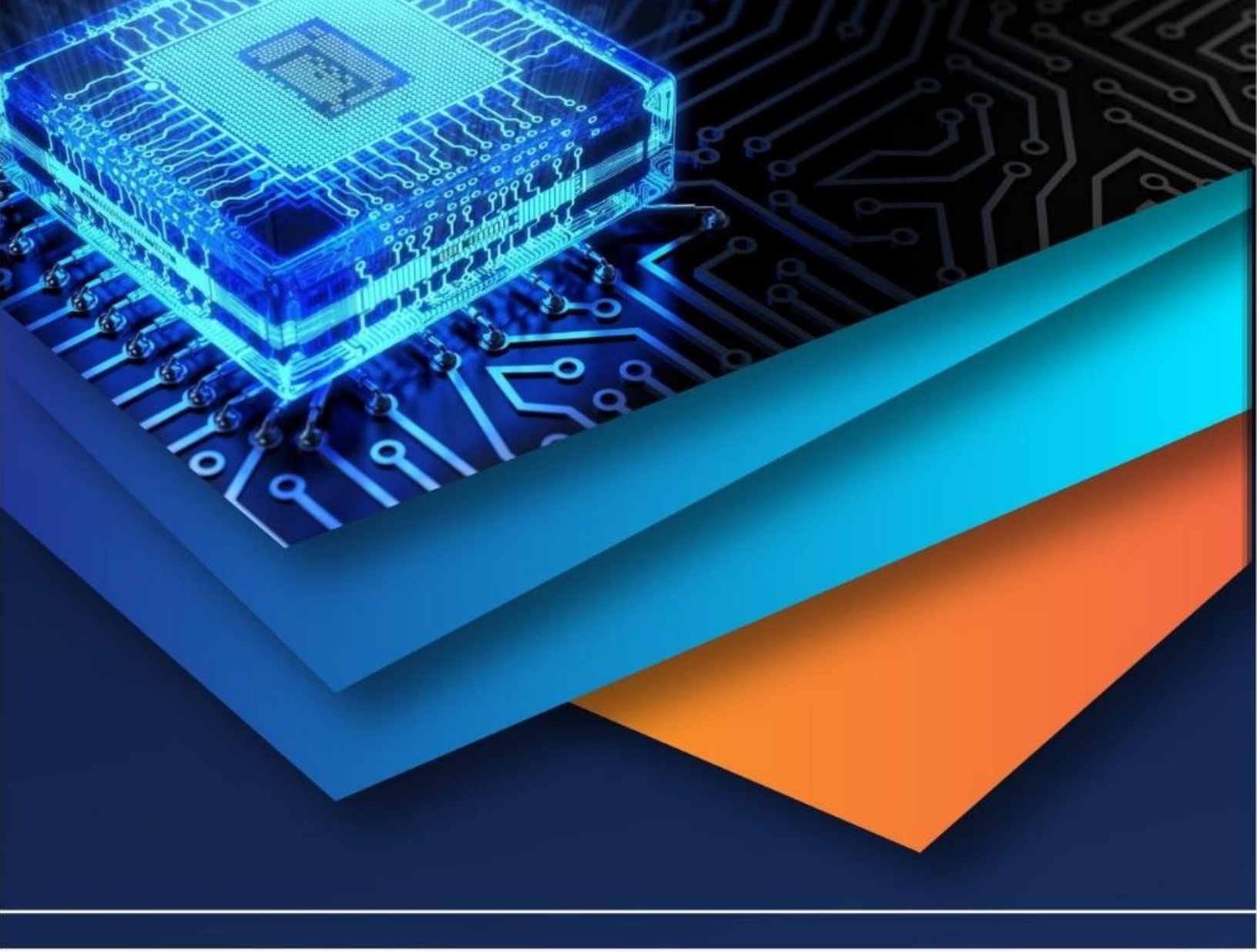
3) Multilingual Expansion: The current model is optimized for English corpora. Future iterations will expand the tokenizer and encoder to support cross-lingual phoneme representations, enabling multilingual synthesis and code-switching.

4) Security and Voice Cloning Safeguards: Given the model's strong zero-shot voice cloning capabilities, future developments must integrate audio watermarking or synthetic speech detection classifiers to prevent misuse and ensure ethical deployment.

REFERENCES

[1] Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang. A review of deep learning based speech synthesis. Applied Sciences, 9(19):4050, 2019.

[2] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561, 2021.
[3] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In International Conference on Machine Learning, pages 5530–5540. PMLR, 2021.
[4] Ye Jia, Heiga Zen, Jonathan Shen, Yu Zhang, and Yonghui Wu. PnG BERT: Augmented BERT on phonemes and graphemes for neural TTS. arXiv preprint arXiv:2103.15060, 2021.
[5] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, Frank K. Soong, Tao Qin, Sheng Zhao, and Tie-Yan Liu. NaturalSpeech: End-to-End Text to Speech Synthesis with Human- Level Quality. arXiv preprint arXiv:2205.04421, 2022.
[6] Yinghao Aaron Li, Cong Han, and Nima Mesgarani. StyleTTS: A Style-Based Generative Model for Natural and Diverse Text-to-Speech Synthesis. arXiv preprint arXiv:2205.15439, 2022.
[7] Yinghao Aaron Li, Cong Han, Xilin Jiang, and Nima Mesgarani. Phoneme-Level BERT for Enhanced Prosody of Text-To- Speech with Grapheme Predictions. In ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.
[8] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. arXiv preprint arXiv:2301.02111, 2023.
[9] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech. In International Conference on Machine Learning, pages 8599–8608. PMLR, 2021.
[10] Rongjie Huang, Max W. Y. Lam, J. Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. FastDiff: A Fast Conditional Diffusion Model for High-Quality Speech Synthesis. In Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI), 2022.
[11] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. Advances in Neural Information Processing Systems, 33:17022–17033, 2020.
[12] Takuhiro Kaneko, Kou Tanaka, Hirokazu Kameoka, and Shogo Seki. iSTFTNet: Fast and Lightweight MelSpectrogram Vocoder Incorporating Inverse Short-Time Fourier Transform. In ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 6207–6211. IEEE, 2022.
[13] Jonathan Shen, Ye Jia, Mike Chrzanowski, Yu Zhang, Isaac Elias, Heiga Zen, and Yonghui Wu. Non-Attentive
[14] Tacotron: Robust and Controllable Neural TTS Synthesis Including Unsupervised Duration Modeling. arXiv preprint arXiv:2010.04301, 2020.
[15] Liu Ziyin, Tilman Hartwig, and Masahito Ueda. Neural Networks Fail to Learn Periodic Functions and How to Fix It. Advances in Neural Information Processing Systems, 33:1583–1594, 2020.
[16] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the Design Space of Diffusion-Based Generative Models



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY