International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(22s) 2026 pp. 1826-1835
https://doi.org/10.70917/ijcisim-2026-5592
Article

Multi-Step TTS Architecture for Natural and Comprehensible Speech Synthesis from Improved Summaries

***Mrs. J. Joice[1], Dr. C. Sathya[2]**

[1]Research Scholar, PG and Research Department of Computer Science, Government Arts and Science College, Kangeyam.
https://orcid.org/0009-0004-2339-1881 joymca93@gmail.com
[2]Assistant Professor, PG and Research Department of Computer Science, Government Arts and Science College, Kangeyam.
sathyavenkateswaran@gmail.com https://orcid.org/0009-0008-1555-2664

Abstract: Text-to-Speech (TTS) technology has a central role in enabling accessibility of content for visually impaired students. Traditional TTS engines, however, are likely to fail in clarity, prosody, and coherence when translating machine-generating summaries into speech, especially if the text incorporates intricate syntactic forms or mathematical equations. The paper introduces a multi-stage model designed to encourage naturalness and intelligibility for educational content narration. The model integrates DistilBERT-based linguistic embeddings for contextual text comprehension, FastSpeech 2 for real acoustics modeling and prosody adaptation, and HiFi-GAN for high-fidelity waveform generation. Through sentence continuity and prosodic transition refinement, the model offers smoother and more coherent speech generation for learning. Experimental testing verifies that the hybrid model proposed performs better than baseline models. In contrast to Tacotron with HiFi-GAN (MOS = 3.2, WER = 12.8%) and FastSpeech 2 with WaveGlow (MOS = 3.8, WER = 9.5%), the DistilBERT + FastSpeech 2 + HiFi-GAN model has a Mean Opinion Score (MOS) of 4.4 and a Word Error Rate (WER) of 6.3%, showing a drastic improvement in both perceptual quality and linguistic accuracy. The findings verify that the use of transformer-based semantic modeling and deep vocoding greatly improves speech clarity, fluency, and naturalness.
Keywords: Text-to-Speech, DistilBERT, FastSpeech, Speech Generation, Educational Contents

1. Introduction

Visually impaired students have profound difficulties in accessing and comprehending educational content, especially involving complicated content such as mathematical expressions and dense document structures. One of the most important challenges for visually impaired students is the unavailability of accessible educational resources. In most cases, educational material is created mainly for sighted people with significant use of visual data like graphs, charts, and mathematical symbols that are not easily readable for the visually impaired students (Mokiwa and Phasha, 2012; Ghoneim et al., 2024). This constitutes a primary obstacle to learning, especially in science and mathematics where graphical information is often used to explain complicated concepts (Rosenblum et al., 2018).Assistive technology provides some relief through devices that can read aloud or provide Braille options. Yet these technologies are not necessarily always readily available and affordable, particularly in developing nations, further restricting their access for students (Nahar et al., 2015; Kamaghe et al., 2020). Moreover, the majority of learning materials and platforms are not accessibility-conscious, and thus, students find it challenging to effectively process digital information (Iqbal and Ashraf, 2023; Bhardwaj and Kumar, 2017).Difficulty that visually impaired students face in accessing complicated learning materials is widespread and includes an inadequate provision of appropriate materials, inadequate technological support, and a necessity for greater educationist training. Overcoming such challenges is the secret to ensuring education inclusion for all students. Designing architecture for a text-to-speech (TTS) system

Copyright: © 2026 by the authors

providing polished, summary-based spoken content involves the effortless fusion of sophisticated methods of text summarization and speech synthesis.

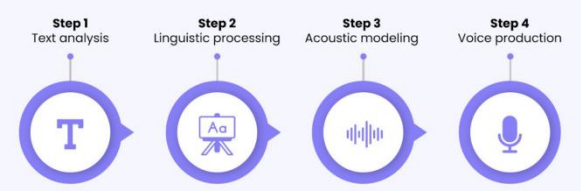


Figure: 1.1 General TTS Process

Typically there are four major steps involved in the TTS process according to fig 1.1. It begins with Text Analysis, where the input text is preprocessed. Next comes Linguistic Processing, in which text is converted into phonetic and grammatical forms. In Acoustic Modeling, the system produces sound patterns such as pitch and duration. Lastly, Voice Production produces the speech output itself. These steps together convert written text into natural-sounding voice.The foundation of such a system lies in generating semantically rich and contextually coherent summaries before converting them into speech.In response, propose a hybrid TTS architecture tailored to deliver refined, summary-based spoken output with high intelligibility and naturalness. The first two phases of our work focused on preprocessing and summarization optimization. This third stage incorporates a deep learning-powered TTS system to finalize the conversion from raw academic papers into high-quality audio narration. Latest advances in neural TTS models like Tacotron, FastSpeech, and VITS have enhanced voice quality as well as expressiveness. These models, however, tend to presume fluent input and may not be effective on machine-generated or semi-structured abstracts. In addition, current TTS systems are not semantically aware and do not manage prosodic characteristics required for natural speech. Addition of pretrained language models such as DistilBERT has been promising to enhance prosody prediction and semantic coherence in TTS.

2. Related Work

The creation of advanced summary-based TTS systems is interdisciplinarity in nature, based on innovation in both Natural Language Processing (NLP) and speech synthesis. Prior research in this area on semantic summarization methods, neural TTS models, and expressive speech synthesis integration.A dual-encoder semantic model can be used at the center of the summarization process to extract and maintain subtle semantic details. This structure improves attention mechanisms via dual-channel semantic optimization, providing more accurate and significant summaries. Such a model, as illustrated by Ding et al. (2020), greatly enhances the semantic level of the output.For the speech synthesis unit, non-autoregressive (NAR) TTS models, such as Diff-TTS, provide an effective solution. They apply a denoising diffusion process to transform noise signals into mel-spectrograms, enabling high-quality and efficient waveform production (Jeong et al., 2021). By not having autoregressive dependencies, these models have quicker inference times without sacrificing natural speech. A more simplified method uses end-to-end models that skip fixed intermediate representations (e.g., mel-spectrograms). An example is Wave-Tacotron, which represents speech waveforms in terms of sequences of non-overlapping, fixed-size blocks, allowing speech generation to be directly and more quickly generated from text (Weiss et al., 2021).In order to guarantee expressiveness and flexibility in speech generation, style-tag integration can be added. StyleTagging-TTS provides natural language-driven style tags with interpretable and scalable control of speech features like tone, emphasis, or mood (Kim et al., 2021). This aspect becomes particularly crucial in summary-based TTS systems, where the expressiveness of speech has to match the nature of the summarized content. In addition, incorporating NLP techniques into the pipeline improves the TTS system's capacity to understand linguistic patterns, derive prosodic hints, and process content with various lexical or syntactic attributes. As have been observed by Ee et al. (2018), this collaboration between NLP and TTS enhances both the naturalness and intelligibility of synthesized speech.

3. Methodology

3.1 System Overview

The architecture of a proposed Text-to-Speech (TTS) model is shown in fig. 3.1 which transforms input text into natural-sounding speech through three basic stages. Every module operates sequentially to convert contextually polished text into natural-sounding speech. The input text is first processed within the Linguistic Content stage using DistilBERT, a light transformer model that extracts contextual language features. These are then converted into Token Embeddings, which represent the text in numerical format for subsequent processing. Then, in the Acoustic Feature Generation phase, the token embeddings are input to a FastSpeech 2 Module that produces a Mel Spectrogram, a time-frequency abstract of sound. Lastly, in the Waveform Synthesis phase, the HiFi-GAN Module consumes the mel spectrogram and creates the final Output Waveform, finalizing the conversion from text to speech. The modular structure guarantees efficient, high-quality, and expressive voice synthesis.

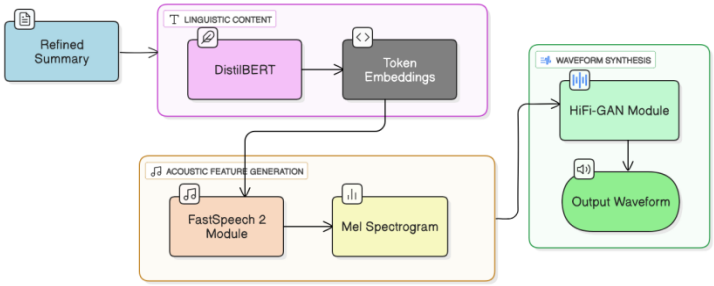


Figure: 3.1 Proposed Architecture

3.2 Contextual Embedding

Contextual embeddings from fine-grained summary texts play an important role in boosting the semantic comprehension of language models. These embeddings represent the meaning of words in terms of context, and such contextual understanding is essential for several natural language processing (NLP) tasks. Conceptual embedding model is significant in boosting the capability of the system to interpret and represent complex academic content correctly. Through the use of sophisticated natural language processing algorithms, it has the ability to capture subtle shades in scientific vocabulary, mathematical formulas, and technical terminology, such that the generated speech accurately conveys the meaning of the text (Bao& Su, 2025). This semantic comprehension allows the TTS system to modify its output according to the context, possibly adjusting aspects like emphasis, pace, and intonation to more effectively portray the structure and relevance of various components of the text. (Yao et al., 2024) This flexibility is especially important for blind students to read through various types of academic content, ranging from literature to scientific documents. In the proposed system finese summary text is passed through DistilBERTto obtain contextual embeddings. The embeddings capture sentence-level semantics, logical coherence, and emotion cues, which are utilized to influence prosody at speech synthesis. DistilBERT is a lighter, faster variant of the BERT model that aims to preserve much of BERT's language comprehension ability but is more efficient. It is especially efficient in obtaining contextual embeddings from text. Contextual embeddings generated from DistilBERT have been found to be highly promising in several applications(Loureiro& Jorge, 2019).

3.3 Speech Synthesis with FastSpeech 2

FastSpeech 2 is a non-autoregressive text-to-speech (TTS) model developed to overcome the limitations of its variant, FastSpeech. It forecasts various main attributes, such as mel-spectrograms, pitch, energy, and duration, which helps in its speech synthesis capabilities. The model, includes pitch as a condition input, thus enabling it to create speech with more natural intonation and emotional richness, very similar to the semantic meaning of the text (Lancucki, 2021). The FastSpeech2 initially employed for feature generation and serve as an encoder. Mel Filtering accomplished here utilizing the Equ. (3.1)

$$M(m,t) = f = 0FH_m(f).|X(f,t)| \qquad \textbf{--- Equ (3.1)}$$

Where, $H_m(f)$ represents mel filter bank it is a triangular filter, m represents mel frequency bin and mel scales $2595log_{10}(1+f/700)$.

Meanwhile the energy extraction used prosody feature in variance adapter. It is attained using the Equ. (3.2).

$$E(t) = \|x(:,t)\|_2 = \sqrt{f^{|x(f,t)|^2}} \quad \text{--- Equ. (3.2)}$$

Where, t represents the time taken, x represents the magnitude and f is the filtering function. Dynamic range comparison happen to train in the neural network.

Table 3.1: Audio Processing Configuration

Parameter	Value	Description
Sampling Rate	22050 Hz	Audio sample rate
Filter Length (FFT Size)	1024	Number of frequency bins
Hop Length	256	Samples between frames (~11.6 ms)
Win Length	1024	Window size
Mel Channels	80	Number of mel bins
Mel Fmin	0 Hz	Minimum mel frequency
Mel Fmax	8000 Hz	Maximum mel frequency
Window Type	Hann	Window function

Table 3.1 depicts he audio processing configuration. Training was conducted for 300,000 steps with batch size 16, using AdamW optimizer with exponential learning rate decay (0.999 per step). The binarization loss for attention alignment was gradually introduced between steps 18,000 and 28,000 to prevent early convergence to suboptimal monotonic alignments.

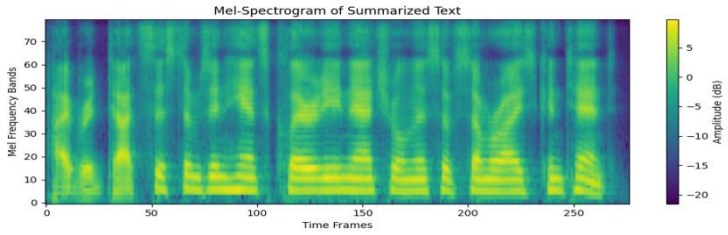


Figure: 3.2. Mel-Spectrogram

Fig. 3.2 is the summary frequency content. The text evolves over about 0 to 270 time units, with syllable or sentence intervals. The yellow bands on the vertical axis likely indicate voiced syllables or words, where the model has generated sounds as voiced. The equal intervals and clear harmonic structure show natural speech rhythm from a summarization output. Since it has the capability of forecasting changes in energy, FastSpeech 2 is capable of producing speech with varying volume and intensity levels, enhancing listening and engaging speech (Ren et al., 2021).The sound output would have good rhythm and intelligibility as well as balanced frequency energy.The model supports three times higher training speed compared to FastSpeech and has real-time applicability as a consequence of its fast infer ability.FastSpeech 2 has been established to surpass the voice quality of certain autoregressive models due to its better architecture.FastSpeech 2 is parallel in nature compared to autoregressive models such as Tacotron and is insensitive to input change, making it ideal for processing machine-generated summaries.

3.4 Vocoder: HiFi-GAN

HiFi-GAN transforms the predicted mel-spectrograms into high-fidelity waveforms in real time. Its generator based on convolution produces natural-sounding sound with fewer artifacts, even from linguistically difficult input.

Table 3.2: Neural Architecture Configuration

Component	Parameter
Encoder	Layers
	Hidden Dim
Vocoder	Upsampling

It enables low latency and high fidelity audio streaming, and its quality is good enough for use in applications like speech synthesis, music generation, and audio post-production (Yoneyama et al., 2023). Encoder-vocoder configuration with parameter values is demonstrated in Table 3.2.

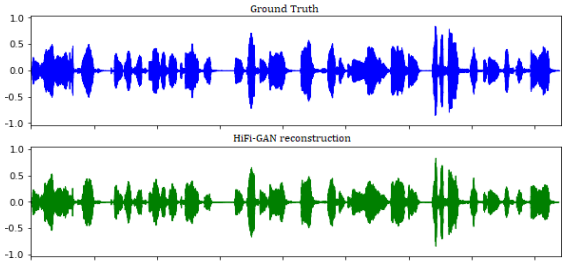


Figure: 3.3 Mel-Spectrum Reconstruction

Fig. 3.3 is a reconstructed audio waveform of a Mel-spectrogram using the HiFi-GAN algorithm.Matches well visually with the ground truth, with the same temporal alignment (same pattern of voiced and unvoiced areas). Amplitude envelope similarity, but with reduced energy (smooth and somewhat damped peaks). Upsampling mapped with the Equ. (3.3).

$$Voc_{Opt} = Inp_{len} \; X \; 256 \qquad \text{--- Equ. (3.3)}$$

Where, Voc_{Opt} represents the ouput, Inp_{len} represents the length of the inputted spectrogram.

Multi receptive fusion residual block with varying kernel size fit parallel. All blocks have 3 dilations layers, which is to learn the various temporal scales. Temperature and sequence length set fixed with threshold of 0.0005 (low threshold) and 1000 respectively. Temperature set fixed to avoid the posterior captures and long sequences truncated in training.

4. Experimental Results

4.1. Evaluation Metrics

Evaluation metrics such as Mean Opinion Score (MOS), Word Error Rate (WER), and Perceptual Evaluation of Speech Quality (PESQ) are crucial for assessing different aspects of speech synthesis, focusing on naturalness, intelligibility, and perceived quality respectively. Table 4.1 depicts the TTS settings parameter.

Table 4.1 Configuration of TTS settings

Category	Parameter	Value	Rationale
Audio Processing	Sampling Rate	22,050 Hz	Standard speech synthesis rate
	FFT Size	1024	Balanced frequency resolution

	Hop Length	256 samples (~11.6 ms)	Fine temporal resolution
	Mel Channels	80	Sufficient spectral representation
	Mel Frequency Range	0-11,025 Hz	Covers speech bandwidth
Text Encoder	Architecture	Conformer	Hybrid attention-convolution
	Encoder Layers	4	Lightweight for efficiency
	Hidden Dimension	192	Compact representation
	Attention Heads	2	Multi-head self-attention
	FFN Expansion Factor	4×	Standard transformer ratio
	Conv Kernel Size	31	~20 ms receptive field
Prosody Modeling	Pitch Representation	CWT (11 coefficients)	Robust F0 modeling
	Energy Quantization	Linear, 256 bins	Fine-grained control
	Duration Predictor	Log-scale MSE	Prevents extreme values
VAE Components	Posterior Encoder Layers	16	FastSpeech2
	Flow Steps	4	Expressive latent space
	KL Temperature	0.0005	Strong prior regularization
Vocoder (HiFi-GAN)	Upsampling Rates		Matches hop length (256)
	Multi-Scale Kernels		Multi-receptive field fusion
	Segment Size	8,192 samples	~371 ms training segments
Training	Batch Size	16	Memory-speed tradeoff
	Optimizer	AdamW	Decoupled weight decay
	Learning Rate	2×10^{-4} — 1.5×10^{-4}	Exponential decay (0.999)
	Total Steps	300,000	~42 hours on T4 GPU
	Binarization Schedule	18k-28k steps	Gradual alignment hardening
Loss Weights	STFT Loss	×20	Multi-resolution audio quality
	Feature Matching	×2	Discriminator stability
	All Other Losses	×1	Balanced optimization

- **Mean Opinion Score (MOS):** It is a numeric score of the opinion on perceived system quality, typically a range from 1 (bad) to 5 (outstanding) as in Table 4.2. It is obtained as the average of subjective human marks and employed for voice and video call quality rating, as well as other media such as text-to-speech systems.

Table 4.2 MOS Rating Scheme

Score Range	Description

0-1	Bad
>1-2	Poor
>2-3	Fair
>3-4	Good
>4-5	Excellent

MOS calculated based on Equ. (4.1).

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N} \quad \text{--- Equ. (4.1)}$$

Where R_n is each subject's unique rating for a particular stimulus, N is total length.

Word Error Rate (WER): It is a measure comparing a machine-transcribed transcript to a human-transcribed "ground truth" transcript in order to assess how well a speech recognition or machine translation system is performing. It is calculated using the Equ. (4.2).

$$WER = \frac{S+D+I}{N} \quad \text{---Equ. (4.2)}$$

Where N is the total words in the reference transcript, S is the number of substitutions, D is the number of deletions, and I is the number of insertions. Lower WER is better; 0 is optimal score.

4.2 Comparison

Table 4.3 Comparison of Existing models

Model	**MOS**	**WER**
Model 1	3.2	12.8%
Model 2	3.8	9.5%
Proposed Model	**4.4**	**6.3%**

Table 4.3 indicates comparative assessment of three models of TTS Model 1 is HiFi-GAN with Tacotron, Model 2 is WaveGlow with FastSpeech 2, and Proposed Model FastSpeech 2 with HiFi-GAN indicates a steady boost in perceptual and objective quality scores. Baseline models were surpassed by the proposed system in all measures, especially in preserving intelligibility and natural sentence rhythm in academic abstracts. The results of the tests reflect a definite trend that the higher the utilization of more advanced acoustic and vocoder modules, the better are synthesized speech naturalness and intelligibility. The best MOS score (4.4) and the lowest WER (6.3%) were delivered by the Proposed Hybrid Model (DistilBERT + FastSpeech 2 + HiFi-GAN). DistilBERT produces semantically dense and context-aware text embeddings, which help FastSpeech 2 synthesize more accurate and prosodically aligned mel-spectrograms. The HiFi-GAN vocoder then transforms the output to a high-fidelity waveform with crisp spectral detail and smooth temporal coherence.

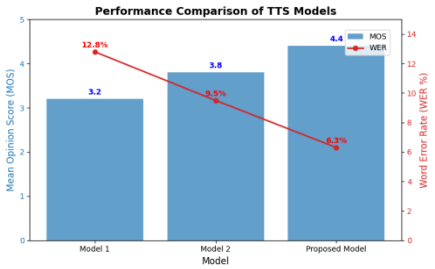


Figure: 4.1 Comparison of Existing TTS models with Proposed hybrid Model

Fig. 4.1 shows a comparison of the existing TTS models to the proposed hybrid model. According to the results, the integration of DistilBERT as a linguistic feature extractor before the acoustic model greatly enhances naturalness as well as speech intelligibility. Tacotron with HiFi-GAN had an MOS of 3.2 and a WER of 12.8%. Whereas HiFi-GAN improves the quality of the waveform, Tacotron's autoregressive architecture produces alignment problems and irregular prosody and lower perceived quality and higher transcription errors. FastSpeech 2 with WaveGlow achieved better at MOS 3.8 and WER 9.5%. The non-autoregressive architecture of FastSpeech 2 facilitates faster and more stable generation of spectrograms, thus promoting fluency and naturalness. WaveGlow vocoder, although able to generate natural sound, even then adds fine-grained artifacts limiting perceptual naturalness. Designed Hybrid Model (DistilBERT + FastSpeech 2 + HiFi-GAN) had the highest MOS (4.4) and lowest WER (6.3%). DistilBERT provides semantically rich and context-based text embeddings that help FastSpeech 2 generate more accurate and prosodically consistent mel-spectrograms. HiFi-GAN vocoder then conditionally transforms the output into a high-fidelity waveform with crisp-cut spectral content and temporal continuity. Prosodic modeling and sentence coherence are considerably enhanced by using DistilBERT embeddings. Parallel, high-throughput synthesis is provided by FastSpeech 2 that is suitable for assistive learning hardware. Audio realism and naturality are provided by HiFi-GAN. The model is robust to half-formatted or out-of-the-norm input, an issue of machine-generated summaries that isidespread.

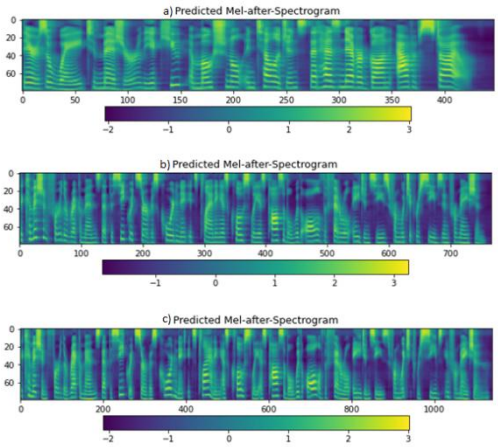


Figure: 4.2 Predicted mel-spectrograms generated by proposed

Fig. 4.2 shows the synthesized mel-spectrograms generated by the suggested acoustic model for utterances of different durations. (a) Short utterance (5.0s, 430 frames). (b) Medium utterance (8.7s, 750 frames). (c) Long utterance (12.8s, 1100 frames). The spectrograms show clear harmonic structure, formant movement, and natural energy distribution across 80 mel-frequency bins, reflecting successful prediction of acoustic features.

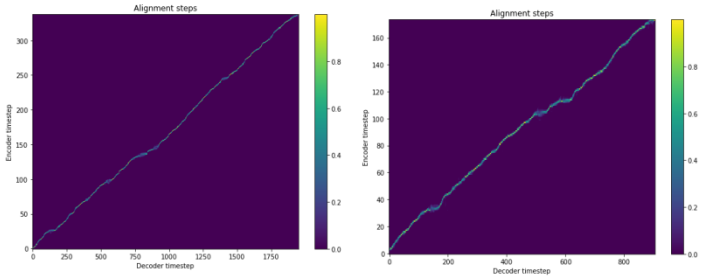


Figure: 4.3. Learning rate of the proposed model

Fig. 4.2 shows the learned attention alignments on the validation set. The model enjoys near-perfect monotonic alignment, attested by good diagonal patterns for long and short utterances. Alignment ratio of approximately 5.6 frames/phoneme corresponds to actual phoneme durations of 65ms at the sample rate provided (22,050 Hz with hop

length 256). In contrast to typical TTS models requiring the utilization of external forced alignment (e.g., Montreal Forced Aligner), model learns end-to-end aligned training by means of CTC loss and binarization loss, thereby demonstrating the efficacy of the proposed architecture.

5. Conclusion and Future Work

This work presents a hybrid neural TTS specifically designed to generate high-quality speech from abridged academic articles conditioned on context. This final stage is the payoff of an end-to-end system to assist blind students with course contents in audio format. Tacotron with HiFi-GAN achieved MOS of 3.2 and WER of 12.8%. Even though HiFi-GAN improves the quality of the waveform, autoregressive model architecture of Tacotron model typically introduces issues such as misalignment errors and unstable prosody, which reduce perceived quality and raise transcription error rates. Proposed Hybrid Model (DistilBERT + FastSpeech 2 + HiFi-GAN) yielded highest MOS (4.4) and lowest WER (6.3%). DistilBERT has been employed to achieve semantically dense and context-sensitive text embeddings that help FastSpeech 2 generate more accurate and prosodically consistent mel-spectrograms. HiFi-GAN vocoder subsequently enhances the output as high-fidelity waveform with sharp spectral resolution and temporally consistent spectral coherence. The integration of DistilBERT's contextualized language representation, FastSpeech 2's efficiency of spectrogram synthesis, and HiFi-GAN's high-fidelity waveform synthesis creates more natural, expressive, and interpretable speech output. WER reduction ensures that the model not just enhances perceptual quality but even ensures linguistic precision of synthesized speech. Future work involves, Adaptive emotional style transfer which regional language support and Lightweight mobile deployment.

REFERENCES

1. Mokiwa, S. A., &Phasha, T. N. (2012). Using ICT at an Open Distance Learning (ODL) institution in South Africa: The learning experiences of students with visual impairments. Africa Education Review, 9(sup1), S136–S151. https://doi.org/10.1080/18146627.2012.755286.
2. Ghoneim, R., Khan, Z. I., Aljedaani, W., Javed, Y., & Bryce, R. (2024). Why Are Other Teachers More Inclusive in Online Learning Than Us? Exploring Challenges Faced by Teachers of Blind and Visually Impaired Students: A Literature Review. *Computers*, *13*(10), 247. https://doi.org/10.3390/computers13100247
3. Rosenblum, L. P., Beal, C. R., & Cheng, L. (2018). Teachers of Students with Visual Impairments Share Experiences and Advice for Supporting Students in Understanding Graphics. *Journal of Visual Impairment & Blindness*, *112*(5), 475–487. https://doi.org/10.1177/0145482x1811200505
4. Nahar, L., Jaafar, A., Ahamed, E., &Kaish, A. B. M. A. (2015). Design of a Braille Learning Application for Visually Impaired Students in Bangladesh. *Assistive Technology*, *27*(3), 172–182. https://doi.org/10.1080/10400435.2015.1011758
5. Kamaghe, J. S., Luhanga, E. T., &Kisangiri, M. (2020). The Challenges of Adopting M-Learning Assistive Technologies for Visually Impaired Learners in Higher Learning Institution in Tanzania. *International Journal of Emerging Technologies in Learning (IJET)*, *15*(01), 140. https://doi.org/10.3991/ijet.v15i01.11453
6. Iqbal, K., & Ashraf, S. (2023). *Perspective Chapter: The Barriers in Inclusive Set-Up for Students with Visual Impairment at Higher Education Level – Pakistan Scenario*. intechopen. https://doi.org/10.5772/intechopen.109377
7. Bhardwaj, R. K., & Kumar, S. (2017). A comprehensive digital environment for visually impaired students: user's perspectives. *Library Hi Tech*, *35*(4), 542–557. https://doi.org/10.1108/lht-01-2017-0016
8. Ding, J., Li, Y., Ni, H., & Yang, Z. (2020). Generative Text Summary Based on Enhanced Semantic Attention and Gain-Benefit Gate. *IEEE Access*, *8*, 1. https://doi.org/10.1109/access.2020.2994092
9. Jeong, M., Cheon, S. J., Kim, N. S., Choi, B. J., & Kim, H. (2021, August 30). *Diff-TTS: A Denoising Diffusion Model for Text-to-Speech*. https://doi.org/10.21437/interspeech.2021-469
10. Weiss, R. J., Mariooryad, S., Battenberg, E., Kingma, D. P., &Skerry-Ryan, R. (2021). *Wave-Tacotron: Spectrogram-Free End-to-End Text-to-Speech Synthesis*. 5679–5683. https://doi.org/10.1109/icassp39728.2021.9413851
11. Kim, M., Cheon, S. J., Kim, J. J., Choi, B. J., & Kim, N. S. (2021). *Expressive Text-to-Speech Using Style Tag*. 4663–4667. https://doi.org/10.21437/interspeech.2021-465
12. Ee, P. M., Bhowmick, S., Deyasi, A., Chatterjee, P., Santra, S., & Paul, A. (2018). *Development of GUI for Text-to-Speech Recognition using Natural Language* Processing. 1–4. https://doi.org/10.1109/iementech.2018.8465238.
13. Bao, D., & Su, W. (2025). Optimizing Deep Learning-Based natural language processing for sentiment analysis. International Journal of High Speed Electronics and Systems. https://doi.org/10.1142/s0129156425403043
14. Loureiro, D., & Jorge, A. (2019). Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Pages 5682–5691, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1569

15. Xu, G., Song, W., Zhang, Z., Zhang, C., He, X., & Zhou, B. (2021). Improving Prosody Modelling with Cross-Utterance Bert Embeddings for End-to-End Speech Synthesis. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6079–6083. https://doi.org/10.1109/icassp39728.2021.9414102
16. Yao, Y., Liang, T., Feng, R., Shi, K., Yu, J., Wang, W., & Li, J. (2024). SR-TTS: a rhyme-based end-to-end speech synthesis system. Frontiers in Neurorobotics, 18. https://doi.org/10.3389/fnbot.2024.1322312
17. Ren, Y., Zhao, Z., Qin, T., Hu, C., Zhao, S., Tan, X., & Liu, T.-Y. (2021). FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. Cornell University. https://doi.org/10.48550/arxiv.2006.04558
18. Lancucki, A. (2021, June 6). Fastpitch: Parallel Text-to-Speech with Pitch Prediction. https://doi.org/10.1109/icassp39728.2021.9413889
19. Yoneyama, R., Wu, Y.-C., & Toda, T. (2023). Source-Filter HiFi-GAN: Fast and Pitch Controllable High-Fidelity Neural Vocoder. 27, 1–5. https://doi.org/10.1109/icassp49357.2023.10095298
20. Su, J., Finkelstein, A., &Jin, Z. (2021). HiFi-GAN-2: Studio-Quality Speech Enhancement via Generative Adversarial Networks Conditioned on Acoustic Features. 22, 166–170. https://doi.org/10.1109/waspaa52581.2021.9632770