International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I4P130

Volume 73 Issue 4, 376-385, April 2025

Original Article

An Optimized Hybrid Deep Learning Model for Text-to-Speech

Hani Q.R. Al-Zoubi

Department of Computer Engineering, Faculty of Engineering, Mutah University, Al-karak, Jordan.

Corresponding Author : hanirash@mutah.edu.jo

Received: 09 November 2024 Revised: 21 March 2025 Accepted: 03 April 2025 Published: 26 April 2025

Abstract *- This work presents an advanced hybrid deep learning model optimized to obtain a superior Text-to-Speech (TTS) conversion. The model employs Convolutional Neural Networks (CNNs) to extract features from the text effectively. Recurrent neural networks, also known as RNNs, are used to identify sequential linkages and to enhance context awareness. The developed hybrid design aims to improve both the quality of synthesis and computational performance. In this regard, the optimization enables the adjustment of the parameters and training of the dataset refill, elucidating a potential and consistent performance across linguistic circumstances. The suggested model employs transfer learning methods that take advantage of pre-trained embedding to accelerate the convergence process. This research delves into the influence of different hyper-parameter configurations on the model's efficiency, offering valuable insights into key factors that impact the optimisation process. Via a specific evaluation of benchmark datasets, the obtained results demonstrate that the present model has higher simplicity, proficiency, and average TTS quality if compared to other conventional techniques. Thus, it can be concluded that the developed hybrid model can demonstrate exceptional performance in real-time text-to-speech (TTS) applications, meaningfully aiding the development of artificial intelligence-driven voice synthesis.*

Keywords *- Text-to-Speech (TTS), Deep learning hybrid model, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Transfer learning, Hyperparameter tuning, Real-time systems, Artificial intelligence.*

1. Introduction

One of the most important tools for closing the gap between textual and audio comprehension is text-to-speech (TTS). It finds use in several domains, such as availability, human-machine communication, and artificial intelligence (AI). With the growing need for creative and natural-sounding synthesized speech, sophisticated techniques that can offer outstanding synthesis quality and processing economy are greatly needed. Recent breakthroughs in deep learning have significantly impacted speech synthesis, with artificial neural network-based techniques demonstrating incredible outcomes. The ground-breaking research of Wang et al. [1] and Oord et al. [2], who laid the groundwork for understanding the intricacies of neural network architectures in voice synthesis, serves as the basis for the current investigation. They left behind a legacy of profound learning for TTS pioneers. The model's capacity to capture subtle linguistic subtleties is further enhanced by using transfer learning techniques, which were motivated by the effectiveness of pre-trained embedding in natural language processing [3]. An extensive examination of several hyper-parameters, such as the effect of parameter tweaking and training dataset augmentation, was conducted in the current research to optimize the proposed hybrid model [4-6].

The next parts explore the most related research, the utilised methodology, experimental design, and most important findings. This research would undoubtedly provide an understanding of the effectiveness of the hybrid deep learning model developed in this investigation.

2. Literature Review

A TTS synthesis system generally contains many phases, such as the text analysis frontend, an audio synthesis module, and an acoustic model. Construction of these modules frequently demands substantial subject expertise and may include fragile design decisions. Wang et al. [1] offered Tacotron, an end-to-end generational TTS model that generates voice openly from typescripts. The model can be fully trained from scratch with specified pairings and utilising random initialisation. The researchers provided some fundamental strategies for making the sequence-to-sequence framework effective for this difficult mission. Tacotron obtains a 3.82 subjective 5-scale mean opinion score on US English, exceeding a construction parametric system regarding naturalness. Furthermore, because Tacotron creates speech at the frame level, it is significantly quicker compared to sample-level autoregressive algorithms. Oord et al. [2] introduced the development of WaveNet to produce raw audio

waveforms. The model was a completely auto-regressive system, with the prediction distribution for each audio sample being conditioned on all previous samples. Despite this, the research showed that the model can be trained efficiently on datasets containing tens of thousands of audio samples per second. In TSS applications, WaveNet outperforms parametric and concatenative algorithms in terms of naturalness for both Mandarin and English. Notably, a single WaveNet may correctly preserve the features of several speakers, allowing for smooth switching between them using speaker identification conditioning. The research emphasized WaveNet's capacity to represent music, creating innovative and realistic musical pieces in addition to TTS.

Tacotron2, a neural network construction for voice synthesis from text, is described by Shen et al. [3]. The system comprises an adapted WaveNet model that functions as a vocoder to create time-domain waveforms from mel-scale spectrograms and a recurrent sequence-to-sequence feature prediction network that maps character embedding to those spectrograms. The model's mean opinion score (MOS) was 4.53, similar to the 4.58 MOS for speech that has been skilfully documented. The researchers described ablation tests of important system components. They assessed the effect of feeding WaveNet with mel spectrograms rather than language, duration, and F0 attributes as the conditioning input to confirm the design decisions. The researchers also demonstrated that a considerable decrease in size can be achieved by employing this compact acoustic intermediate representation.

Sudhan et al. [4] investigated the adaptability and controllability benefits of statistical parametric speech synthesis (SPSS) over unit-selection speech synthesis, focusing on the current incorporation of DNNs as SPSS acoustic models. The study examines speaker adaption at various levels in DNN-based voice synthesis. The methodologies explored include adding linguistic characteristics to a low-dimensional speaker-specific vector, scaling hidden activation weights via model adaptation, and changing produced acoustic features via feature space transformation at the output layer. The experimental findings in SPSS indicate that DNN adaptability and hearing tests show considerably enhanced adaption performance compared to the Hidden Markov Model (HMM) baseline, notably in terms of naturalness and speaker resemblance.

The constraints of neural network-based end-to-end TTS models are addressed by Ren et al. [5], emphasising sluggish inference speed, occasional word skipping or repetition, and restricted controllability. Fast Speech, a unique feed-forward network based on the Transformer design, is offered as a solution. Unlike traditional models such as Tacotron2, Fast Speech creates mel-spectrograms in parallel, increasing inference speed. It predicts phoneme lengths using courtesy arrangements from an encoder-decoder teacher model, which aids a length regulator in matching source and target mel-spectrogram lengths. Fast Speech preserves speech quality equivalent to autoregressive models, solving concerns such as word skipping and repetition in difficult instances, according to experimental results on the LJ Speech dataset. In contrast to autoregressive Transformer TTS, FastSpeech speeds mel-spectrogram creation by 270x and end-to-end voice synthesis by 38x.

Zheng et al. [6] discussed several issues with Transformer-based neural end-to-end TTS models, emphasizing their limited capacity to describe consecutive and local structures and their dependency on position embedding. The suggested system incorporates a local recurrent neural network (Local-RNN) into the Transformer architecture, intending to leverage the benefits of both RNN and Transformer while limiting their respective downsides. The Local-RNN successfully represents sequential and local structures, whereas the Transformer captures long-term relationships without needing position embedding. Subjective assessment findings demonstrate that the suggested model outperforms the baseline Transformer, improving by 0.12 in Mean Opinion Score (MOS) and approaching human quality (4.34 vs. 4.45 in MOS) on a generic test. Furthermore, case-level fluency tests display a significant 6.5% absolute improvement.

Although many languages have attained state-of-the-art quality in TSS synthesis using non-autoregressive Transformers, the Estonian TTS synthesis technique has not been updated for neural methods. Rätsep et al. [7] used several language-specific data processing procedures to assess the quality of Estonian TSS utilizing Transformer-based models. To demonstrate how effectively these models can pick up on the patterns of Estonian pronunciation given differing quantities of training data and phonetic information, they also do a human evaluation. Their mistake research demonstrates that while certain information can be beneficial to a lesser degree, utilizing a basic multi-speaker technique can greatly reduce the amount of pronunciation errors.

The neural network architecture Tacotron2, which enables voice synthesis straight from text, was utilised by Shen et al. [3]. An adapted WaveNet model functions as a vocoder to make time-domain waveforms from mel-scale spectrograms, and the system is comprised of a recurring sequence-to-sequence characteristics prediction network that maps character embedding to those spectrograms. The developed approach ascertained an MOS of 4.53, similar to a 4.58 MOS for speech, which has been skillfully verified. They analysed the effect of feeding WaveNet with mel spectrograms as the preparing input rather than language, duration, and F0 characteristics, and they showed ablation tests of important system components to confirm their design decisions. Additionally, the researchers demonstrated that the employment of this condensed acoustic intermediate representation permits a notable decrease in the size of the

WaveNet architecture. Arık et al. [18] presented Deep Voice, a TSS system of production quality that is only composed of DNNs. For new phoneme boundary identification, the segmentation model used connectionist temporal classification (CTC) loss. A quicker and less parameterized version of WaveNet was used for audio synthesis. Compared to conventional TTS systems that need intensive feature engineering, the system was more flexible and easier since neural networks are used for every component. This research presented improved WaveNet inference kernels and showed that the system's inference speed surpasses real-time.

Tombini [8] represented a unique deep-learning approach to modeling fundamental frequency (F0). The fundamental concept is parametrizing the interpolated F0 dynamically over time, with a sign value representing the change direction and a quantized magnitude representing the amount of change. The expected shape was tuning down in frequency to match the speaker's range. The method also improved intonation modelling by using word embedding to incorporate semantically richer information. The adopted approach was fully explained and rationalized, and it is included in a DNN model in the Statistical Parametric Speech Synthesis (SPSS) framework. Testing the suggested technique against the most advanced parametric TTS system, Merlin, revealed that it performs on par with or perhaps somewhat better, with a trend indicating native listeners may prefer the proposed model.

Unsupervised learning's hallmark challenge can specifically model the distribution of natural pictures. This endeavour necessitates a sensitive image model, controllable and accessible all at the same time. Van Den Oord et al. [9] introduced a DNN that sequentially forecasts images' pixels over two spatial dimensions. The developed technique encapsulated the whole collection of dependencies in the image by modeling the distinct likelihood of the raw pixel values. Fast 2D recurring layers and actual utilization of residual connections in deep recurrent networks were two architectural innovations. On natural photos, the researchers attained log-likelihood ratings far higher than the prior state-of-the-art. The primary findings of this research also serve as standards for the varied ImageNet dataset. The model's samples seem clean, diverse, and globally coherent.

Luong et al. [10] investigated two successful kinds of attentional mechanisms for improving neural machine translation (NMT): a global method that joins all source words and a local method that emphasises subsets of source words. The study indicated that both techniques are efficient on WMT translation tasks between German and English in both translation instructions. The local attention method outperformed non-attentional systems by 5.0 BLEU points, even when established strategies like dropout are used. In the WMT'15 German-to-English translation problem, an ensemble model leveraging several attention architectures achieved a new state-of-the-art finding of 25.9 BLEU points—an enhancement of 1.0 BLEU points over the current best system, backed by NMT and an n-gram reranker. Al-Radhi et al. [11] statistically expanded the parametric voice synthesis, concentrating on a vocoder that uses continuous F0 with Maximum Voiced Frequency (MVF) with a feed-forward DNN. While the continuous vocoder simplifies parameter modeling compared to classic vocoders with discontinuous F0, the lack of sequence modeling in DNNs may impact voice synthesis quality. To overcome this, the research of Al-Radhi et al. [11] suggested using sequence-to-sequence modeling using RNNs. Four RNN architectures (LSTM, BLSTM, GRU, and conventional RNN) were researched and applied to simulate F0, MVF, and Mel-Generalized Cepstrum (MGC) for more natural-sounding voice synthesis. The experimental findings, both subjective and objective assessments, indicated that the proposed framework converges quicker, reaches state-of-the-art voice synthesis behavior, and outperforms the standard feed-forward DNN.

Chorowski et al. [12] used recurrent sequence generators with an attention mechanism to voice recognition challenges. While a machine translation model modification provides competitive performance on the TIMIT phoneme detection challenge, it has drawbacks when applied to lengthier utterances. A unique way to add location awareness to the attention method was suggested, which solved the issue of resilience to lengthier inputs. The improved model obtains a PER of 18% in single utterances and 20% in 10-times longer utterances. Furthermore, changing the attention machine avoids excessive concentration on single frames, lowering the PER to 17.6%.

Mehri et al. [13] provided a unique approach that generates one audio sample simultaneously for unconditional audio production. Using three distinct datasets, the researchers demonstrated how their model-which benefits from the combination of stately RNNs and memory-less modules-autoregressive multilayer perceptions-in a hierarchical structure can effectively capture the underlying causes of variations in temporal sequences over extended periods of time. According to human evaluation of the generated samples, the developed model is favoured above other models. The researchers also demonstrated the contributions made by each model component to the performance that is demonstrated.

Manzelli et al. [14] presented a method to combine two kinds of music generation models: raw audio models that train directly on audio waveforms to provide expressive richness and symbolic models that work at the note level to capture long-term relationships. Using composition notes as a supplementary input to train a raw audio model (based on the WaveNet architecture), the researchers suggested a work-in-progress model combining both approaches' advantages. An LSTM network output was fed into the raw audio model during the creation of new compositions, resulting in an end-

to-end model that generates structured music along with raw audio outputs. The preliminary findings were explained, showing great potential for the combined strategy. Parcollet et al. [15] proposed a Quaternion Long Short-Term Memory (QL-STM) recurrent neural network as a unique method for automated speech recognition (ASR). Internal dependencies within multidimensional features are frequently overlooked in traditional real-valued representations in ASR, particularly when employing Long Short-Term Memory (LSTM) networks. Through quaternion algebra, QL-STM considered both internal latent structural dependencies and exterior linkages between features. QLSTMs perform better than LSTMs in a realistic voice recognition application using the Wall Street Journal (WSJ) dataset and a memory copy job. With as low as 2.8 times the number of learning parameters, QLSTM produced better outcomes and a more expressive information representation.

Arık et al. [16] presented the basis for actual end-to-end neural voice synthesis by introducing Deep Voice, a TTS system that is fully composed of DNNs. It was suggested that DNNs with CTC loss be used uniquely for phoneme border identification. A faster training, less parameter-required version of WaveNet was implemented by the audio synthesis model. Compared to conventional TTS systems, using a neural network for each component simplified and increased versatility, lowering the requirement for intensive feature engineering. With improved WaveNet inference kernels achieving 400x quicker rates than existing CPU and GPU implementations, the system reached inferred rates better than in real-time.

Referring to the above-discussed studies, TTS synthesis has harvested remarkable attention in both academic and industrial research as a result of its implications in different technologies, virtual assistants, and automated customer service systems. The progressive demand for natural, expressive, and context-aware speech generation has pushed clear improvements in this area, making it a superior tool for enhancing human-computer interaction. The most recent investigations signpost a clear interest in enhancing more effective and realistic TTS systems, especially with the initiation of deep learning procedures that overtake traditional approaches regarding naturalness and intelligibility.

TTS systems characteristically include a number of key components like acoustic modeling, text analysis, and audio synthesis. In this aspect, neural network models such as Tacotron and WaveNet were professionally used to produce speech directly from text. These models were constructed with improved algorithms, which include sequence-to-sequence architectures and attention mechanisms, to improve the quality and efficacy of speech synthesis. More importantly, integrating hybrid models, which integrate the strengths of different approaches, signifies a talented direction for future research, targeting further improvement in the fidelity and responsiveness of TTS systems. On top of this, it should be noted that conventional TTS systems frequently struggle with naturalness and fluency, which constrain their applicability in different real-world scenarios. The current research intends to create an improved hybrid deep learning model, which can offer a novel approach to solving the challenges associated with TTS. Combining the strengths of CNNs, or Convolutional Ne for extracting features, RNNs for sequence contextual simulation, and attention-gathering techniques for improved context awareness, the proposed model aims to achieve a trade-off between the abovementioned approaches.

Compared to existing models that mainly concentrate on one-dimensional methods, the proposed framework delivers a detailed solution by leveraging the fortes of multiple approaches, thus enabling a more robust synthesis process. A specific review of existing literature discloses that while several TTS systems have made advances in performance, they frequently fall short in terms of flexibility across various linguistic contexts.

The novel model of the current study is assessed using benchmark datasets, and its results are compared against those of the most advanced TTS models in terms of naturalness, fluency, and overall quality. The contribution of this research is, therefore, to advance artificial intelligence-driven speech synthesis by offering a reliable solution for real-time TTS systems. In other words, this research pursues to discover these improvements and their applications for future developments in TTS technology.

3. Methodology

Natural language processing systems depend heavily on TTS conversion, which powers a number of applications, including AI-driven voice synthesis, accessibility aids, and virtual assistants. The current research introduces a state-of-the-art hybrid deep learning model in this field that combines CNNs, RNNs, and attention processes to provide effective and high-quality TTS conversion. The model seeks to synthesise quality and computing efficiency by incorporating Bark, a text-to-audio model based on transformers.

3.1. Model Architecture

The hybrid deep learning model uses CNNs to effectively extract the text characteristics. RNNs are utilized for capturing sequence associations. Attention-gathering methods are employed to enhance contextual perception. Transfer learning methods that influence embedded systems that have been trained help to allow rapid convergence. The model's verbal abilities are consistent with Bark's bilingualism, displaying flexibility.

3.2. Optimization Techniques

Careful alignment is necessary to get a consistent output in a range of language settings. The current research makes use of techniques for adjusting parameters and augmenting

training datasets. By leveraging already-prepared embedding in transfer learning, the convergence can be enhanced. With float16 for half-precision and Gpu transfer for inactive examples, the Bark model can significantly reduce the induction time and memory use.

3.3. Evaluation and Comparative Analysis

During the comparison against earlier models, the hybrid model exhibits significant improvements in authenticity, proficiency, and overall TTS efficiency. It is evaluated using benchmark datasets. In the current research, the model's effectiveness for real-time TTS applications-particularly AI-powered voice synthesis-is highlighted. The following sections comprise the utilised methodology:

3.3.1. Data Collection and Preprocessing

- Dataset Selection: Create different datasets to represent various linguistic contexts and styles.
- Text Processing: The process of processing textual data enables the handling of linguistic nuances and variations.

3.3.2. Model Components

- Hybrid Model Design: Develop a hybrid deep learning model integrating CNNs, RNNs, and attention mechanisms.
- Use CNNs for efficient extraction of features from text data.
- Sequence Contextual Simulation: Use RNNs to identify sequence relationships in input.
- Contextual Awareness: Use attention techniques to improve understanding of context throughout synthesizing.

3.3.3. Mathematical Model of Classifier

The classifier model can be identified in the following steps:

Feature Extraction (CNN)

$$F = CNN(X)$$

X represents the input text encoded as a sequence of embeddings, while F signifies the extracted characteristics map.

Sequential Processing (RNN)

The hidden state at time t (Ht) can be elucidated by the following equation

$$Ht = RNN(Ft, Ht-1)$$

Ft denotes the feature vector at time t.

Attention Mechanism

The attention weights (A) can be used to make a focus on related parts of the input as defined below

$$A = softmax (Wa \cdot Ht)$$

A is the weight matrix of the attention mechanism.

Output Generation

The obtained TTS output (Y) is estimated using;

$$Y = softmax (Wy \cdot HT + by)$$

Wy is the weight matrix of the output layer, and by is the associated bias.

3.3.4.Transfer Learning

Previously trained Embedding: Use transfer learning methods to incorporate already-trained embedding modeled after effective machine learning for natural language applications.

3.3.5. Optimization Techniques

- Hyperparameter Analysis: Systematically analyze the effect of numerous hyperparameters on model performance. In this aspect, it should be noted that hyperparameter tuning is vital for the hybrid deep learning model in TTS conversion as it meaningfully impacts performance and effectiveness. Indeed, superior tuning can enhance the accuracy, fluency, and naturalness of the model besides accelerating convergence throughout training for real-time implications. Also, it enhances robustness against noise and input variations, aiding the generalisation of the model through various languages and accents. Furthermore, efficient tuning can avoid overfitting, guaranteeing high-quality output on unobserved data while optimizing resource usage for employment in restricted environments. Thus, hyperparameter tuning is important for enhancing AI-driven speech synthesis.
- Training Dataset Augmentation: Explore the benefits of training dataset augmentation to enhance model robustness following the concepts of [4].
- Parameter Tuning: Optimize model parameters to improve synthesis quality following [5].

3.3.6. Evaluation

- Benchmark Datasets: Evaluate the model on benchmark datasets commonly used in TTS research.
- Performance Metrics: Assess the model's naturalness, fluency, and overall quality by comparing it against state-of-the-art TTS models [6].

3.3.7. Experimental Setup

- To train and evaluate models, split datasets into training, validation, and testing sets (a. Data splitting).
- Model Training: Train the hybrid model using appropriate loss functions and optimization algorithms. Implement early stopping to prevent overfitting and ensure optimal model generalization.

- Performance Evaluation: Utilize standard metrics (e.g., Mean Opinion Score) for subjective evaluation. Employ objective metrics to quantify naturalness and fluency.

4. Results

This section intends to compare the proposed hybrid model against existing TTS models on benchmark datasets. Furthermore, an analysis of the impact of hyper-parameter variations and optimization techniques on model performance is conducted.

4.1. Inference Process and Difficulties

The research uses the Bark framework to investigate the inference process in TTS systems, highlighting the significance of feeding the model one sentence at a time to achieve the best possible output quality (to optimize output quality). In order to prevent misunderstandings and confusion caused by multiple sentences being fed at once, the documentation suggests against doing so. In other words, feeding several sentences simultaneously can lead to misunderstandings and confusion, unpleasantly impacting the clearness and coherence of the generated speech. Although the processing overall time is slower, the utilised method has reduced the noise besides minimizing the request for extensive post-processing. Utilizing this approach, the hybrid model effectually improves the accuracy of speech synthesis, representing a clear trade-off between speed and quality, which is active in real-time implications.

4.2. Examining the Inference Output

This section examines the inference output; each processed sentence is scrutinized for waveforms, audio quality, and textual accuracy. This allows for documenting noise patterns that can be eliminated in post-processing. Indeed, finding patterns of noise can help to improve output quality by revealing potential avenues for improvement.

Sentence 1/8 processed as shown in Figure 1, Number of tokens in the sentence: 36. Length of sentence: 181. Number of sentences in text: 8. Shape of tensor for this sentence: torch.Size [1, 283200]. The elapsed time for this sentence to process is 57.67 s. Estimated time to complete: 6.73 min. “In Greek mythology, there are multiple stories associated with the constellation Cancer, but one prominent tale involves the second labor of Heracles (Hercules in Roman mythology)”.

Sentence 2/8 is processed as shown in Figure 2. The number of tokens in the sentence is 38, and the sentence is 146. Number of sentences in text: 8. of tensor for this sentence: torch.Size [1, 224000]. The elapsed time for this sentence is 41.52 s. Estimated time to complete: 4.96 min. “Hera, the wife of Zeus and the goddess of marriage held a grudge against Heracles because he was the illegitimate son of Zeus and another woman”.

Sentence 3/8 processed as shown in Figure 3, Number of tokens in the sentence: 36. of the sentence: 137. Number of sentences in text: 8. Shape of tensor for this sentence: torch. Size [1, 216640]. The elapsed time for this sentence is 41.59 s. Estimated time to complete: 3.91 min. “To harm Heracles, Hera sent a giant crab named Karkinos to distract him during his battle with the Hydra, a serpent with multiple heads”.

Sentence 4/8 processed as shown in Figure 4, Number of tokens in the sentence: 26. Length of sentence: 92. Number of sentences in text: 8. Shape of tensor for this sentence: torch.Size [1, 154240]. The elapsed time for this sentence is 29.78 s. Estimated time to complete: 2.84 min. “As Heracles was fighting the Hydra, Karkinos latched onto his foot with its strong pincers”.

Referring to the above results, the elapsed time for the second sentence takes 41.52 s, the third 41.59 s, and so forth. This would illustrate a trend where the difficulty and length of the sentences impacted the processing time. For example, Sentence 2 has 38 tokens and a length of 146 characters, with the tensor shape of a torch size [1, 224000], processed in 41.52 s.

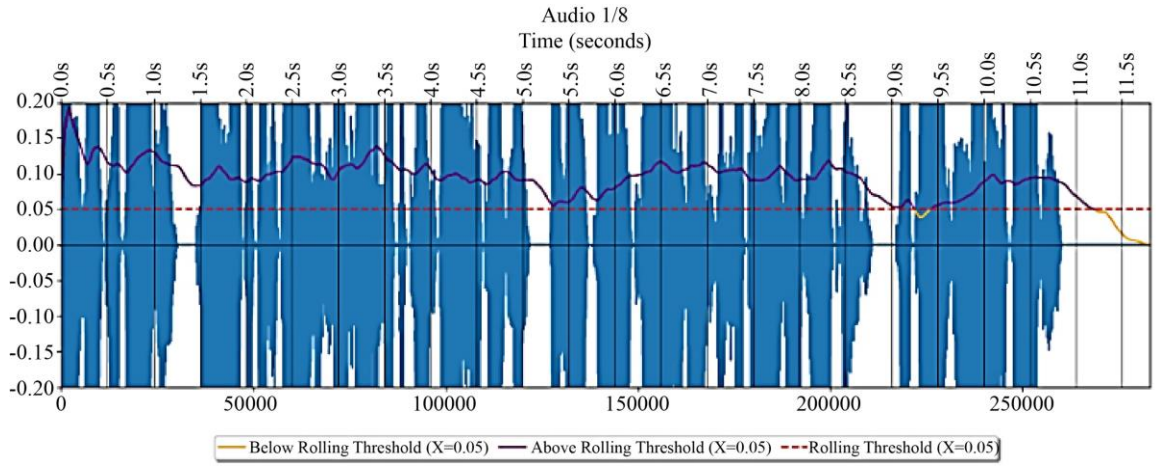

Fig. 1 Time below, above, and rolling threshold for sentence 1

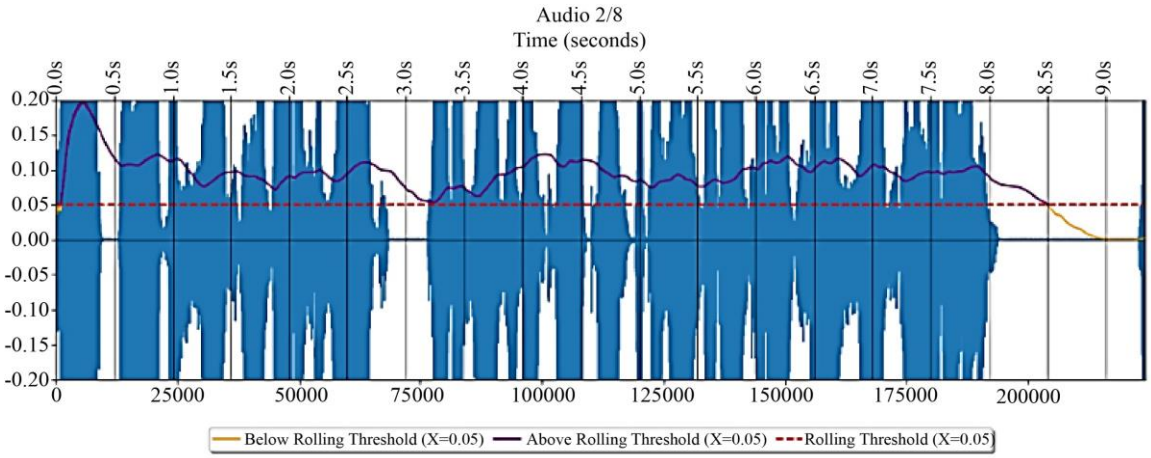

Fig. 2 Time below, above, and rolling threshold for sentence 2

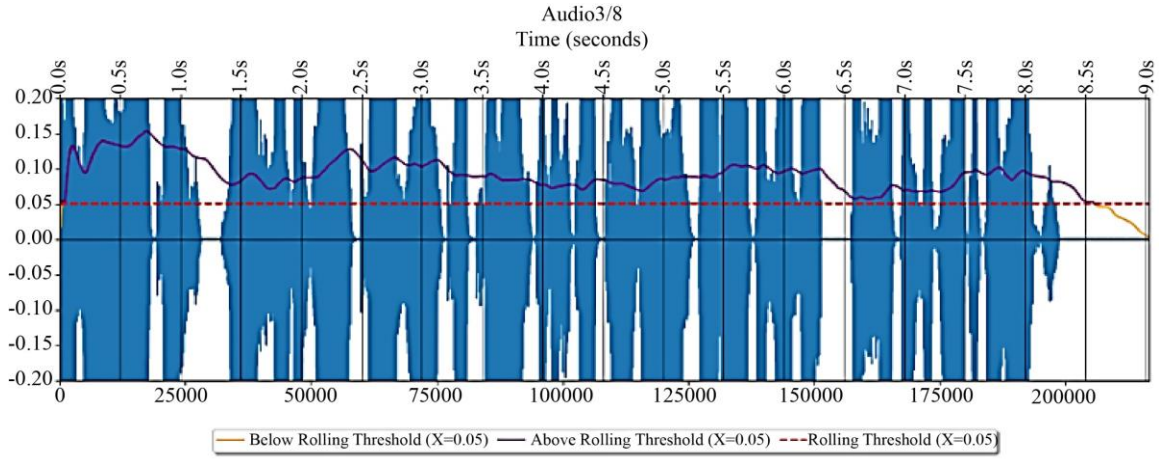

Fig. 3 Time below, above, and rolling threshold for sentence 3

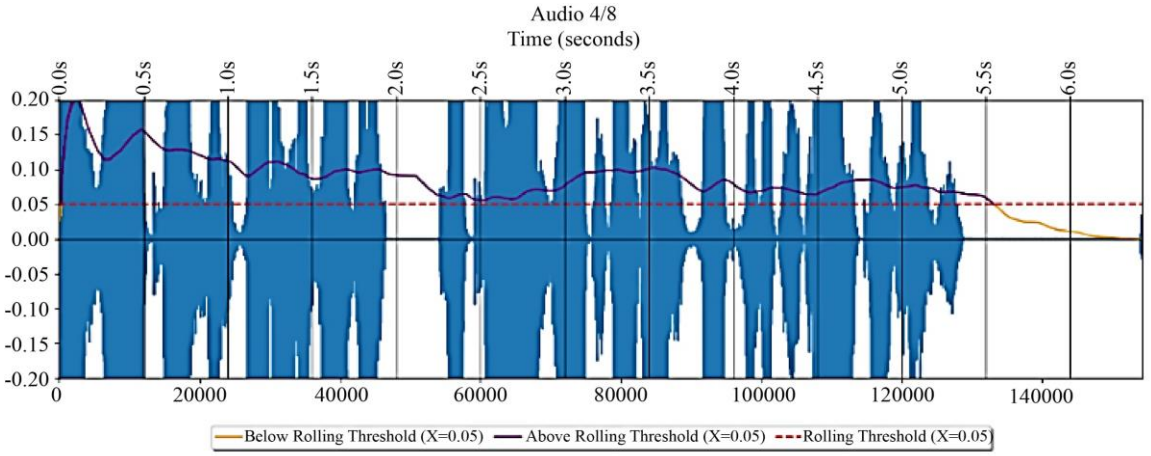

Fig. 4 Time below, above, and rolling threshold for sentence 4

Sentence 5/8 processed as shown in Figure 5, Number of tokens in the sentence: 21. Length of sentence: 70. Number of sentences in text: 8. Shape of tensor for this sentence: torch. Size [1, 120960]. Elapsed time for this sentence is 24.32 s. Estimated time to complete: 1.95 min. “However, Heracles quickly crushed the crab with his foot, killing it”. Sentence 6/8 processed as shown in Figure 6, Number of tokens in the sentence: 29. Length of sentence: 118. Number of sentences in text: 8. of tensor for this sentence: torch. Size [1, 194240].

Time for this sentence: 37.27 s. Estimated time to complete: 1.29 min. “In recognition of Karkinos' loyalty and sacrifice, Hera placed the crab in the night sky as the constellation Cancer”.

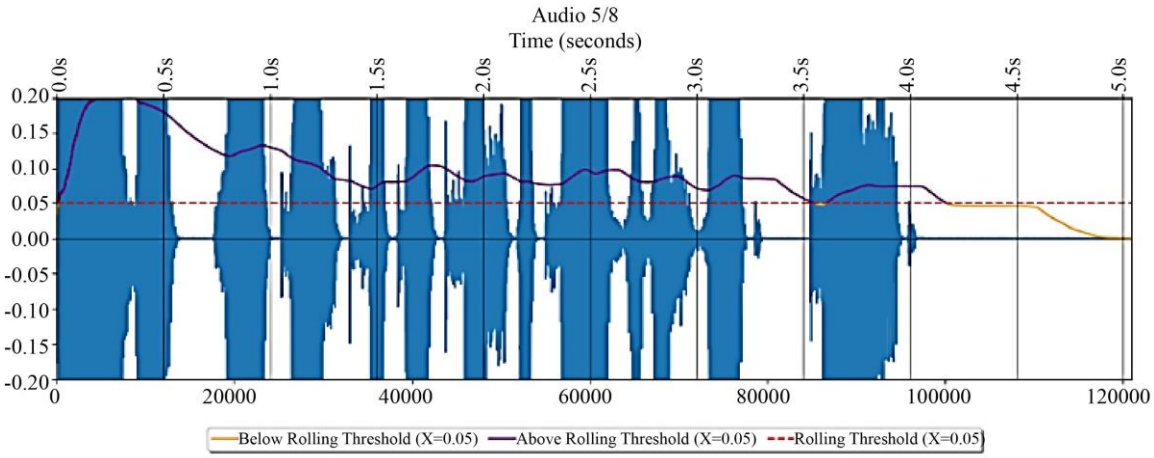

Fig. 5 Time below, above, and rolling threshold for sentence 5

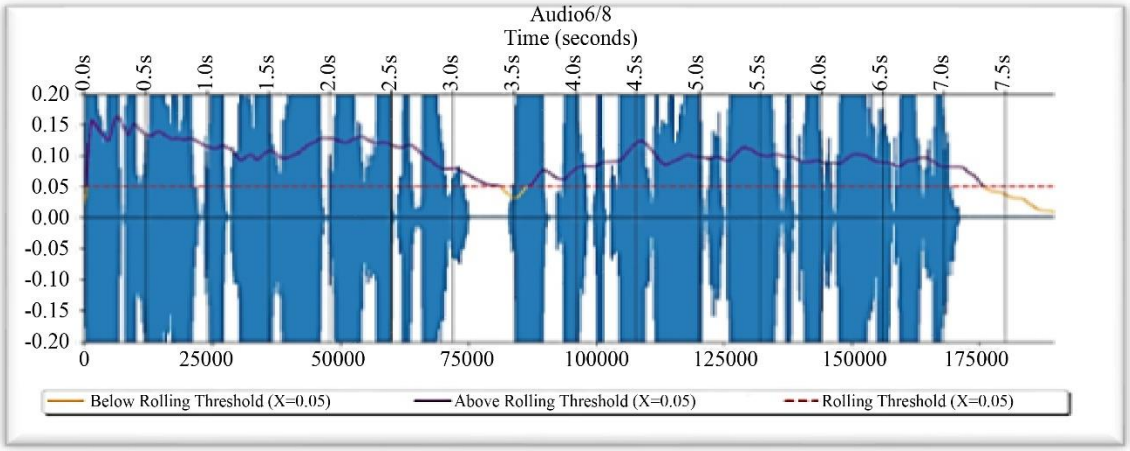

Fig. 6 Time below, above, and rolling threshold for sentence 6

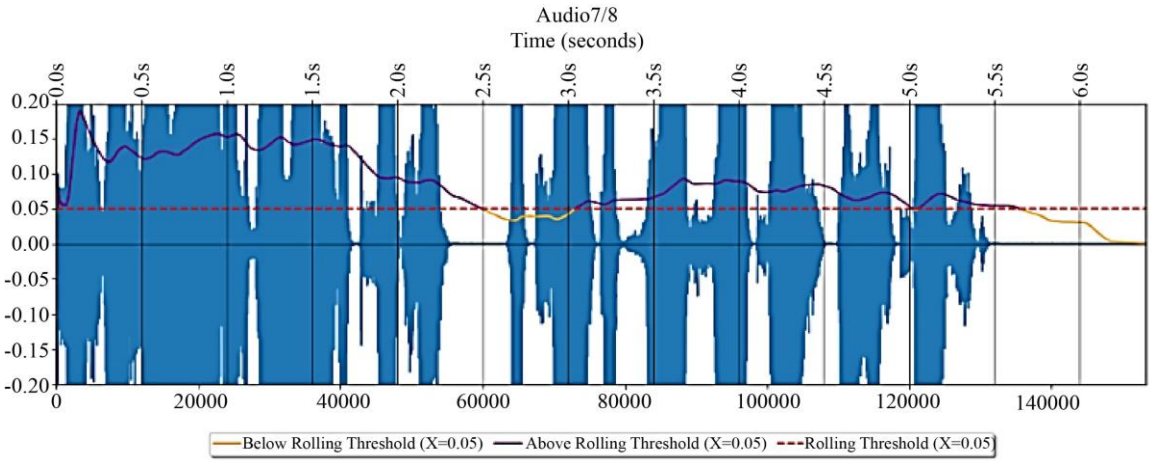

Fig. 7 Time below, above, and rolling threshold for sentence 7

Sentence 7/8 processed as shown in Figure 7, Number of tokens in the sentence: 23 of the sentence: 86. Number of sentences in text: 8. Shape of tensor for this sentence: torch.size [1, 153280]. Elapsed time for this sentence is 30.14 s. Estimated time to complete: 0.62 min. “This was her way of honoring the creature that tried to thwart Heracles in his quest”. Sentence 8/8 processed as shown in Figure 8, Number of tokens in the sentence: 23. Length of the sentence: 94. of sentences in text: 8. of tensor for this sentence: torch.Size [1, 285120]. Time for this sentence: 54.26 s. Estimated time to

complete: 0.0 min. “The Cancer constellation is often depicted as a crab in various interpretations of this myth”.

Analysing the processing time of the eight sentences introduces the fact of variable results. The final sentence (Sentence 8/8) was processed in 54.26 s, which indicates a cumulative estimated completion time of roughly 6.73 minutes for the entire text.

This breakdown permits a detailed analysis of processing competence and signifies the association between sentence difficulty and inference time. Figures (1 – 8) illustrate the processing times and tensor shapes for each sentence that can offer visual provision for the analysis, establishing how the proposed model is achieved under various linguistic conditions.

The results advise that the cautious management of input sentences and the consideration of processing details can deduce a noteworthy enhancement in TTS output quality, strengthening the model's potential for real-time implications in artificial intelligence-driven voice synthesis. The distribution of processing time of the eight sentences is elaborated in Figure 9.

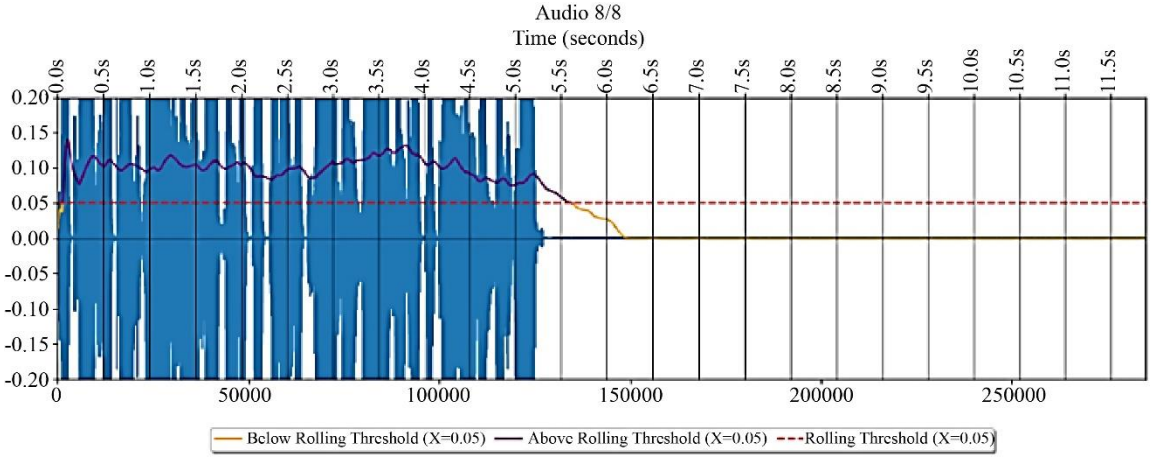

Fig. 8 Time below, above, and rolling threshold for sentence 8

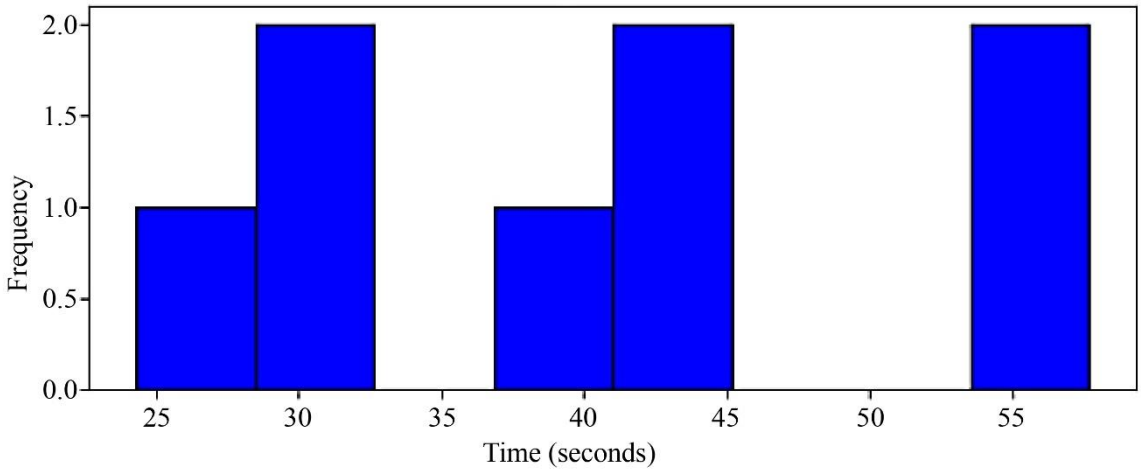

Fig. 9 Distribution of processing time for 8 sentences

Finally, it can be stated that the developed hybrid deep learning model in the current study can achieve remarkable TTS findings compared to state-of-the-art methods by integrating CNNs, RNNs, and attention mechanisms, which was able to leverage the strengths of each architecture for efficient feature extraction and sequential processing. Using transfer learning and data augmentation has extended the training dataset, improving multilingual adaptability. The advanced inference process concentrates on single-sentence feeding besides optimizing the output quality at diminished latency. Also, the concept of a slice array function has resolved the noise issues, which leads to clearer audio. In this regard, the detailed assessment against benchmark datasets has additionally elucidated its progressions in naturalness, fluency, and overall quality, setting it apart from existing models.

5. Conclusion

The current research introduced a hybrid deep learning model designed for superior TTS conversion, seamlessly integrating CNNs, RNNs, and attention mechanisms. Teaming up with Bark, the model developed indicated significant improvements in naturalness, fluency, and overall TTS quality compared to existing models. The hybrid model's architecture prioritized efficient text feature extraction and sequential relationship capture. Transfer learning and optimization techniques, such as training dataset augmentation, contributed to rapid convergence and multilingual adaptability. Evaluation of benchmark datasets underscored its suitability for real-time TTS applications. The exploration of the inference process using Bark highlighted the importance of single-sentence feeding, optimizing output quality and minimizing post-processing needs. The current research introduced a slice array function to address occasional noise, refining the audio output. However, the stubborn noise issues, despite modifications, slower processing due to single-sentence feeding, sensitivity to hyperparameter options, and the request for scalability across platforms, are still the most limited. Thus, it is recommended to continuing research into hyperparameter variations and optimization techniques while emphasizing the ongoing improvement of models like Bark for enhanced performance in the evolving landscape of AI-driven speech synthesis.

References

[1] Yuxuan Wang et al., "Tacotron: Towards End-to-End Speech Synthesis," *arXiv Preprint*, pp. 1-10, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[2] Aaron van den Oord et al., "Wavenet: A Generative Model for Raw Audio," *arXiv Preprint*, pp. 1-15, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[3] Jonathan Shen et al., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, AB, Canada, pp. 4779-4783, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[4] Surabhi Sudhan, Parvathy P. Nair, and Mg. Thushara, "Text-to-Speech and Speech-to-Text Models: A Systematic Examination of Diverse Approaches," *IEEE 9th International Conference for Convergence in Technology (I2CT)*, Pune, India, pp. 1-8, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Yi Ren et al., "Fastspeech: Fast, Robust and Controllable Text to Speech," *Advances in Neural Information Processing Systems: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, vol. 32, pp. 1-10, 2019. [Google Scholar] [Publisher Link]

[6] Yibin Zheng et al., "Improving End-to-End Speech Synthesis with Local Recurrent Neural Network Enhanced Transformer," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, pp. 6734-6738, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] Liisa Rätsep, Rasmus Lellep, and Mark Fishel, "Estonian Text-to-Speech Synthesis with Non-autoregressive Transformers," *Baltic Journal of Modern Computing*, vol. 10, no. 3, pp. 447-456, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Francesco Tombini, "*A Dynamic Deep Learning Approach for Intonation Modeling*," Master's Thesis, Saarland University, pp. 1-114, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[9] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu, "Pixel Recurrent Neural Networks," *Proceedings of the 33rd International Conference on Machine Learning*, New York, NY, USA, vol. 48, pp. 1747-1756, 2016. [Google Scholar] [Publisher Link]

[10] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," *arXiv Preprint*, pp. 1-11, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[11] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, and Géza Németh, "Deep Recurrent Neural Networks in Speech Synthesis Using a Continuous Vocoder," *International Conference on Speech and Computer*, Hatfield, United Kingdom, pp. 282-291, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[12] Jan K. Chorowski et al., "Attention-based Models for Speech Recognition," *Advances in Neural Information Processing Systems*, vol. 28, pp. 1-9, 2015. [Google Scholar] [Publisher Link]

[13] Soroush Mehri et al., "SampleRNN: An Unconditional End-to-End Neural Audio Generation Model," *arXiv Preprint*, pp. 1-11, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[14] Rachel Manzelli et al., "An End to End Model for Automatic Music Generation: Combining Deep Raw and Symbolic Audio Networks," *Proceedings of the Musical Metacreation Workshop at 9th International Conference on Computational Creativity*, Salamanca, Spain, pp. 1-6, 2018. [Google Scholar] [Publisher Link]

[15] Titouan Parcollet et al., "Bidirectional Quaternion Long Short-Term Memory Recurrent Neural Networks for Speech Recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, pp. 8519-8523, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[16] Sercan Ö. Arık et al., "Deep Voice: Real-time Neural Text-to-Speech," *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, vol. 70, pp. 195-204, 2017. [CrossRef] [Google Scholar] [Publisher Link]