iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

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

DOI: https://doi.org/10.22214/ijraset.2025.70467

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 13 Issue V May 2025- Available at www.ijraset.com

Research Paper on Text to Audio Converter using NLP

Anant Sirohi[1], Akshay Yadav[2], Vani Rastogi[3]
Dept of Computer Science MIET, Meerut, India

Abstract: The development of text-to-speech (TTS) systems has advanced significantly with the introduc on of deep learning-based models. This paper inves gates the impact of various deep learning architectures, such as WaveNet and Tacotron 2, on the naturalness of synthesized speech. By leveraging convolu onal neural networks (CNNs) and recurrent neural networks (RNNs), we explore techniques for improving prosody, intona on, and speech quality. Our experiments show that the integra on of a en on mechanisms and vocoder models leads to more accurate and human-like speech output, par cularly in complex sentence structures. Addi onally, we examine the poten al of TTS systems in mul lingual and emo onal speech synthesis, showing promising results in genera ng speech with diverse accents and emo ons.
Keywords: Text-to-Speech, Deep Learning, Wave Net, Tacotron, Speech Synthesis, Mul lingual TTS, Emo onal Speech.

I. INTRODUCTION

Text-to-speech (TTS) systems aim to convert wri en text into spoken language. While early TTS systems relied on concatenate methods that s tched together pre-recorded speech units, recent advancements have shi ed towards neural network-based models. These models, especially those u lizing deep learning architectures, have greatly improved the naturalness and expressiveness of synthesized speech. In this paper, we explore the evolu on of neural TTS systems, focusing on models such as WaveNet (developed by DeepMind) and Tacotron 2 (developed by Google). These models rely on large-scale deep neural networks that are capable of genera ng high-quality, natural-sounding speech by learning from vast amounts of speech data.

A. Background

The first significant breakthrough in TTS research came with the advent of sta s cal parametric speech synthesis, which used hidden Markov models (HMMs) to model the speech waveform. However, the limita ons of this approach, including unnatural prosody and robo c-sounding voices, led to the development of more sophis cated methods. With the success of deep learning, models such as WaveNet and Tacotron have shown remarkable improvements in genera ng realis c speech by learning directly from raw audio data.
WaveNet was one of the first models to generate audio directly from waveform data, which bypassed the need for tradi onal signal processing techniques. However, while it produced highly realis c audio, it was computa onally expensive. To address this, Tacotron 2 introduced a more efficient pipeline, where text is first converted into a spectrogram (a visual representa on of sound), and then converted into audio using a vocoder.

B. Research Objec ves

This paper aims to:

1) Inves gate the role of deep learning models in improving TTS naturalness.
2) Analyze the performance of Tacotron 2 and WaveNet models in terms of audio quality, prosody, and real- me synthesis.
3) Explore the poten al of TTS systems for mul lingual speech synthesis and emo onal tone genera on.

II. RELATED WORK

Many TTS systems have emerged over the years, with notable early systems like Fes val and MBROLA. However, the introduc on of deep learning-based systems has marked a turning point in speech synthesis.
In Wave Net, Oord et al. (2016) introduced a model that generates raw audio waveforms using a deep neural network, producing highly realis c speech. Since then, many varia ons of WaveNet have been explored to make the system more efficient and suitable for real- me applications. Another breakthrough came with Tacotron, which u lized sequence-to-sequence models to convert text into a spectrogram. Wang et al. (2017) introduced Tacotron 2, which combined a sequence-to-sequence network for text-to-spectrogram conversion with a WaveNet vocoder for high-quality waveform genera on. This two-step process dras cally improved the naturalness and intelligibility of synthesized speech.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

A. Mul lingual TTS

Recent research has focused on expanding TTS capabili es to support mul ple languages. Jia et al. (2018) proposed a mul lingual TTS system that learns a shared representa on for mul ple languages, making it possible to generate natural-sounding speech in various languages without needing separate models for each language.

III. METHODOLOGY

To evaluate the effec veness of deep learning models for TTS, we implemented both Tacotron 2 and WaveNet using standard datasets, such as LJSpeech (a single-speaker dataset) and VCTK (a mul lingual dataset). The models were trained on NVIDIA V100 GPUs, and we u lized TensorFlow for model implementa on.

A. Tacotron 2 Architecture

Tacotron 2 consists of two main components:

1) Encoder: The text input is first tokenized and then passed through an encoder that converts it into a sequence of phoneme representa ons.
2) Decoder: The decoder predicts a spectrogram from the encoded phonemes. We use a WaveNet vocoder to convert this spectrogram into a waveform.
3) WaveNet Architecture: The Wave Net model generates raw audio directly from the input, and its architecture is based on dilated convolu ons. For this experiment, we trained a mul -speaker Wave Net model using the LJSpeech dataset.

IV. BLOCK DIAGRAM

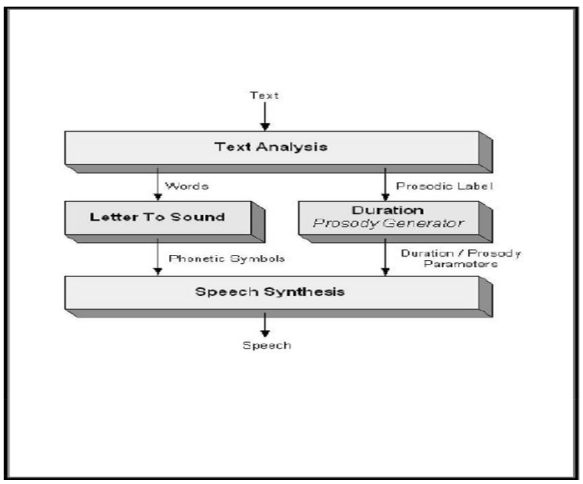


Fig. 1

V. FLOW DIAGRAM

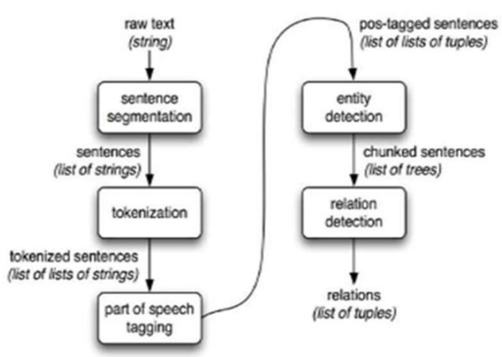


Fig. 2

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Raw text: This is the input text that needs to be processed.
Sentence segmenta on: This step divides the text into sentences.
Tokeniza on: This step splits each sentence into individual words or tokens.
Part of speech tagging: This step iden fies the gramma cal category of each token (e.g., noun, verb, adjec ve).
En ty detec on: This step iden fies named en es in the text (e.g., people, organiza ons, loca ons).
Rela on detecon: This step Idenfies relaonships betweenen es in the text.
Chunked sentences: This step groups tokens into phrases to be er understand the meaning of the text.
Pos-tagged sentences: This step assigns gramma cal tags (e.g., "noun", "verb") to each token in the sentence.

VI. RESULTS

1) A er running a command on we get a URL, we got this screen :-

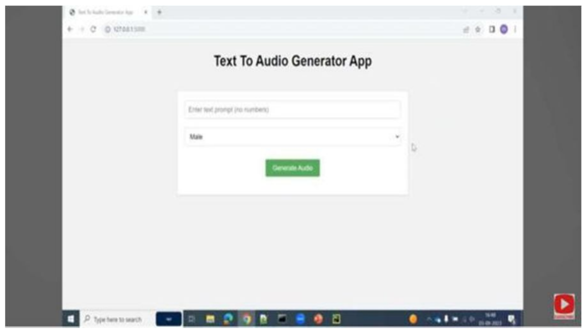


Fig. 3

2) Now we will enter the text that we want to convert into audio

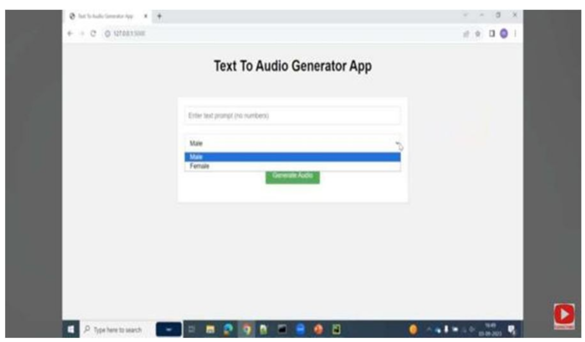


Fig. 4

3) Now we have to select the gender of voice.

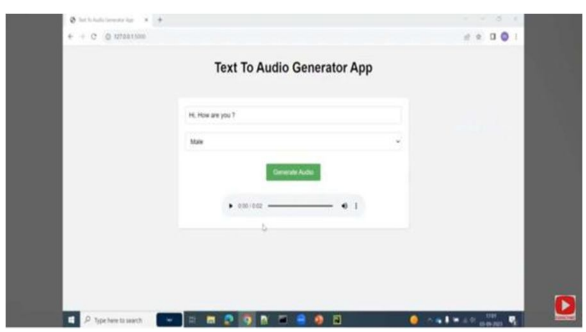


Fig. 5

4) Finally, we will get the audio conversion of the text.

©IJRASET: All Rights are Reserved

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

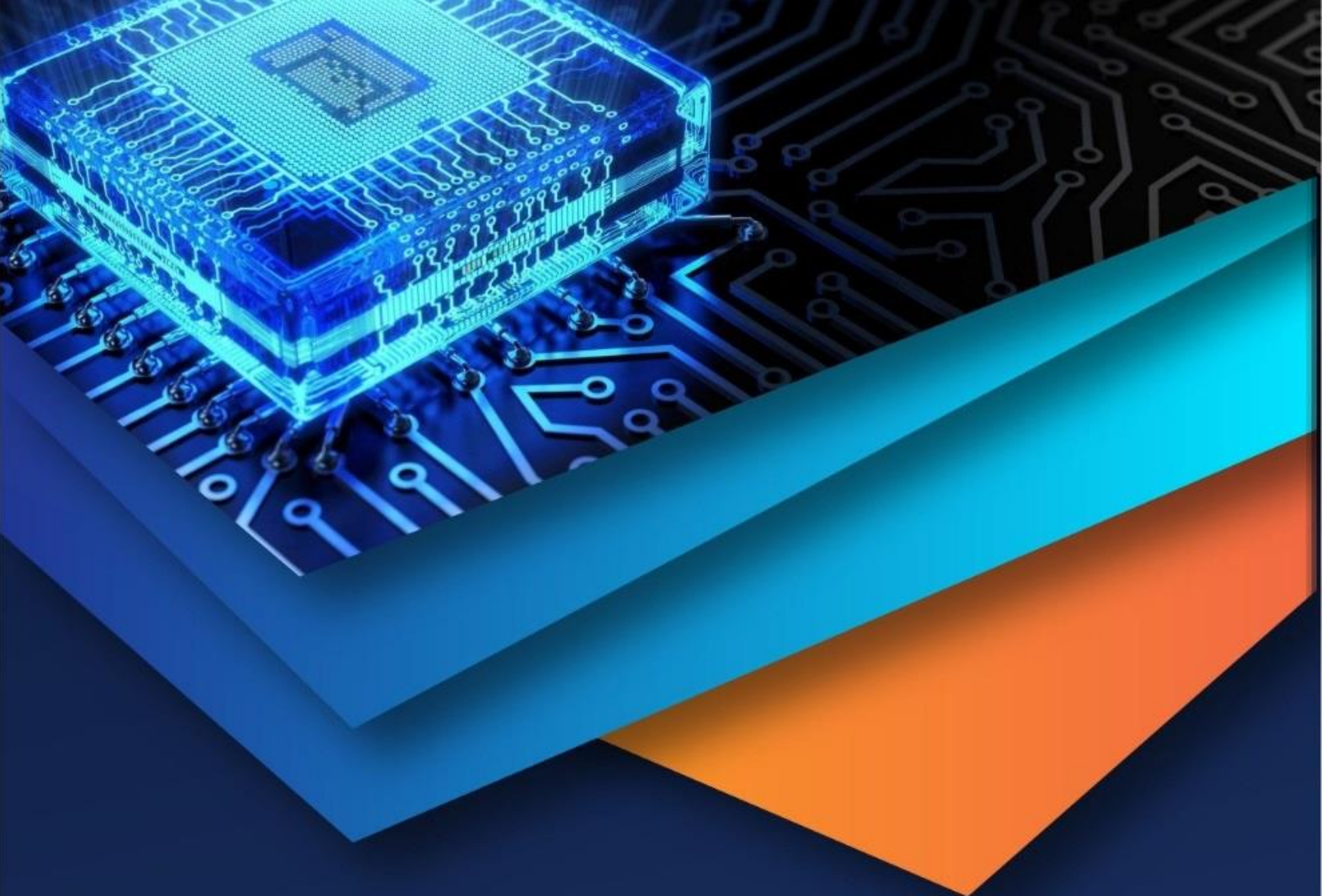
VII. CONCLUSION

This research shows that deep learning-based TTS models, especially WaveNet and Tacotron 2, are capable of producing highly natural-sounding speech. Furthermore, advancements in mul lingual and emo onal speech synthesis highlight the poten al for TTS systems to be applied in a wide range of applica ons, from virtual assistants to audiobook narra on.

Future work will focus on improving real-me synthesis capabili es and exploring the use of emo on modeling in TTS systems to further enhance the expressiveness of generated speech.

REFERENCES

[1] Shen, J., et al. (2018). "Tacotron 2: Genera ng Human-like Speech from Text." Proceedings of the 35th Interna onal Conference on Machine Learning arXiv:1609.03499.
[2] Ren, Y., et al. (2019). "FastSpeech: Fast, Robust, and Controllable Text to Speech." arXiv:1905.09263.
[3] Ping, W., et al. (2018). "ClariNet: Parallel Wave Genera on in End-to-End Text-to-Speech." arXiv:1807.07281.
[4] Kim, J., et al. (2020). "Mellotron: Towards Real-Time Expressive Speech Synthesis with Tacotron." arXiv:2004.04452.
[5] Yang, J., et al. (2020). "Mul -Speaker Mul -Language Speech Synthesis with Tacotron." Proceedings of Interspeech 2020.
[6] Kim, S., et al. (2020). "Parallel WaveGAN: A fast waveform genera on model based on genera ve adversarial networks." IEEE Interna onal Conference on Acous cs, Speech, and Signal Processing (ICASSP 2020).
[7] Liu, Y., et al. (2020). "HiFi-GAN: Genera ve Adversarial Networks for Efficient and High-Quality Speech Synthesis." arXiv: 2010.05646.
[8] Liu, C., et al. (2020). "Voice Transformer Network: A Deep Neural Network for Text-to-Speech with Mul –Speaker and Emotional Varia on." IEEE Transacons on Audio, Speech, and Language Processing.



10.22214/IJRASET

INDEX COPERNICUS
45.98

ISRA JIF
IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)