International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. **11** Special Issue **03** | **2026**

Text to Speech System

Omkar Parab[1], Shubham Mote[2], Vishranti Gawas[3], Rohan Gawade[4], Pratik Rane[5], Heenali Korgoankar[6]

[12.3.4]*Student, Electronic and telecommunication, Metropolitan institute of technology and management, Sindhudurg, Maharashtra, India*
[5,6]*Asst. Professor, Electronic and telecommunication, Metropolitan institute of technology and management, Sindhudurg, Maharashtra, India*

DOI: 10.5281/zenodo.20609765

ABSTRACT

A Text-to-Speech (TTS) system converts written text into spoken words using speech synthesis technology. This project presents the design and implementation of a low-cost embedded Text-to-Speech system using the ESP32 developed by Espressif Systems. The ESP32 microcontroller provides powerful processing capability along with built-in Wi-Fi and Bluetooth, making it suitable for Internet of Things (IoT) and embedded voice applications. In this system, text input is processed by the ESP32 and converted into audio signals using a speech synthesis method or a connected Text-to-Speech library. The generated speech output is then played through a speaker connected to the microcontroller. The system is designed to be compact, cost-effective, and easy to integrate with various smart devices. This Text-to-Speech system can be used in applications such as voice notifications, assistive devices for visually impaired users, smart home systems, and automated information systems. The proposed approach demonstrates how embedded platforms like the ESP32 can be effectively used to implement real-time speech generation in modern electronic systems.

Keyword: - ESP32 - Text-to-Speech (TTS) - Speech Synthesis - Embedded Systems - Internet of Things (IoT)

1.INTRODUCTION

A Text-to-Speech (TTS) system is a technology that converts written text into spoken audio. It allows electronic devices to communicate information to users through human-like speech. TTS systems are widely used in smart assistants, accessibility devices for visually impaired people, navigation systems, and automated announcements. In recent years, microcontrollers have made it possible to build compact and low-cost speech systems. One such powerful microcontroller is the ESP32 developed by Espressif Systems. The ESP32 is widely used in embedded and Internet of Things (IoT) applications because it offers built-in Wi-Fi, Bluetooth connectivity, high processing capability, and low power consumption.

In a Text-to-Speech system using ESP32, the microcontroller receives text input from a user or a connected device. The text is then processed and converted into speech using a speech synthesis algorithm or by accessing a cloud-based TTS service. After processing, the generated audio signal is sent to a speaker through an audio amplifier or digital-to-analog converter. The ESP32 platform is suitable for implementing lightweight speech applications such as voice notifications, talking devices, smart home announcements, and educational tools. By integrating the ESP32 with speakers and software libraries, developers can create efficient embedded Text-to-Speech systems capable of producing understandable speech output. Overall, using the ESP32 for a Text-to-Speech system provides a cost-effective, compact, and flexible solution for embedded voice applications in modern electronics and IoT systems.

2.LITERATURE REVIEW

Text-to-Speech (TTS) technology has been widely researched as a method for converting written text into spoken language using speech synthesis techniques. Early TTS systems were developed using rule-based methods that relied on predefined pronunciation rules and phoneme generation. These systems produced understandable speech but often sounded robotic and lacked natural intonation. With the advancement of digital signal processing and artificial intelligence, modern TTS systems have significantly improved in terms of speech quality and naturalness. Neural network-based models such as WaveNet developed by Google DeepMind have enabled the generation of highly realistic and natural speech by learning from large speech datasets. These approaches use deep learning techniques to model speech patterns, pronunciation, and prosody more accurately.

In the field of embedded systems, researchers have focused on implementing lightweight TTS systems on microcontrollers and IoT devices. The ESP32 developed by Espressif Systems has become a popular platform

for such applications due to its high processing power, integrated Wi-Fi and Bluetooth connectivity, and low power consumption. Several studies demonstrate that the ESP32 can efficiently handle audio processing tasks and interact with cloud-based TTS services. Recent research has explored combining ESP32 with external modules, cloud APIs, or pre-recorded phoneme libraries to generate speech output. These systems are commonly used in smart home devices, voice alert systems, assistive technology for visually impaired individuals, and automated announcement systems. Overall, the literature indicates a shift from traditional rule-based TTS systems toward AI-based speech synthesis and embedded implementations. The use of ESP32 enables the development of compact, cost-effective, and network-enabled Text-to-Speech systems suitable for modern IoT applications.

3. PROPOSED SYSTEM ARCHITECTURE AND HARDWARE DESIGN

1. Proposed System Architecture

The proposed Text-to-Speech (TTS) system is designed to convert input text into audible speech using an embedded microcontroller platform. The system uses the ESP32, developed by Espressif Systems, which provides high processing capability along with built-in Wi-Fi and Bluetooth connectivity.
The architecture consists of several functional modules that work together to generate speech from text input.

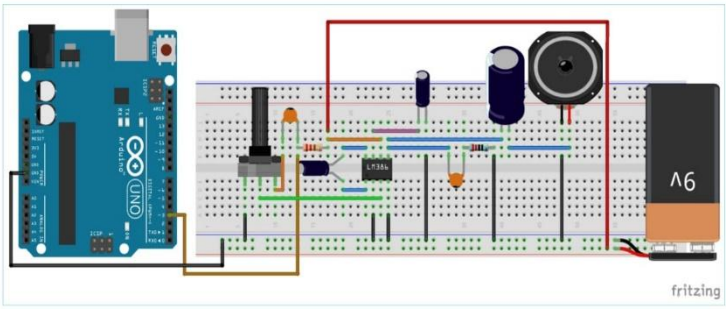

Fig - 1: Schematic diagram of the text to speech system

1.1 Input Module

The input module is responsible for receiving text data. The text can be provided through:
Serial communication from a computer Mobile application via Wi-Fi or Bluetooth Stored messages in memory

1.2 Text Processing Module

In this stage, the input text is analyzed and prepared for speech synthesis. The process includes: Text normalization (handling numbers, abbreviations, and symbols)
Tokenization (splitting text into words) Phoneme conversion for correct pronunciation

1.3 Speech Synthesis Module

The processed text is converted into speech signals using a speech synthesis algorithm or pre-recorded phoneme library. The ESP32 processes the text and generates digital audio data representing the spoken output.

1.4 Audio Output Module

The generated audio signal is sent to an external audio amplifier and speaker. This module converts digital audio signals into audible sound that can be heard by the user.

1.5 Communication Module

The ESP32 provides built-in wireless connectivity that allows communication with external devices such as smartphones or computers. This enables remote control and real-time text input for speech generation.
System Flow
Text Input → Text Processing → Speech Synthesis → Audio Amplifier → Speaker Output

2. Hardware Design

The hardware design of the proposed system includes several components that work together to implement the Text-to-Speech functionality.

2.1 ESP32 Microcontroller

The core component of the system is the ESP32. It performs text processing, speech generation, and communication with external devices.

2.2 Speaker

A speaker is used to output the generated speech. It converts electrical audio signals into audible sound.

2.3 Audio Amplifier Module
An audio amplifier is used to increase the strength of the audio signal before sending it to the speaker for clearer sound output.

2.4 Power Supply
The system requires a stable power supply, typically 5V or 3.3V, to operate the ESP32 and connected modules.

2.5 Communication Interface
Wi-Fi and Bluetooth features of the ESP32 enable communication with other devices such as smartphones or computers for sending text input.

4. METHODOLOGY AND EXPERIMENTAL SETUP

3.1 Methodology
The methodology of the proposed Text-to-Speech (TTS) system describes the steps followed to convert written text into spoken audio using an embedded microcontroller platform. The system uses the ESP32 developed by Espressif Systems to process the text and generate speech output.

Step 1: Text Input
The system begins by receiving text input from a user. The text can be entered through:
Serial communication from a computer
A mobile application using Wi-Fi or Bluetooth Pre-stored text messages in the system memory

Step 2: Text Processing
The ESP32 processes the input text to prepare it for speech generation. This step includes:
Text normalization (handling numbers, symbols, and abbreviations) Breaking sentences into words
Converting words into phonemes for proper pronunciation

Step 3: Speech Synthesis
In this stage, the processed text is converted into digital speech signals. The ESP32 runs a speech synthesis algorithm or uses stored phoneme audio samples to generate understandable speech.

Step 4: Audio Signal Generation
The digital audio signals produced by the ESP32 are sent to an audio output interface. These signals represent the synthesized voice.

Step 5: Speech Output
The audio signals are amplified using an audio amplifier and played through a speaker, allowing the user to hear the generated speech.

3.2 Experimental Setup
The experimental setup defines the hardware and software environment used to implement and test the Text-to-Speech system.

3.2.1 Hardware Components
The system consists of the following hardware components:
ESP32 Microcontroller: Main processing unit responsible for text processing and speech generation. Speaker: Used to produce the audible speech output.
Audio Amplifier Module: Amplifies the audio signal for clear sound. Power Supply: Provides stable voltage (3.3V or 5V) to operate the system.
Communication Interface: Wi-Fi or Bluetooth for sending text input to the ESP32.

3.2.2 Software Environment
The following software tools are used for system development: Programming Language: C/C++
Development Platform: Arduino IDE or ESP-IDF Operating System: Windows or Linux
Libraries: ESP32 audio and communication libraries for speech processing.

3.2.3 Testing Procedure
Upload the TTS program to the ESP32 using the development environment.
Send sample text input to the ESP32 through serial monitor or wireless communication. The ESP32 processes the text and generates speech signals.
The output audio is played through the connected speaker. The speech quality and clarity are evaluated.

3.2.4 Performance Evaluation
The system performance is evaluated based on: Clarity of generated speech

Accuracy of pronunciation Response time of speech generation
Stability of the system during operation

5. DESIGN ANALYSIS AND EXPECTED OUTCOMES

1. Design Analysis

The design of the Text-to-Speech (TTS) system focuses on creating an efficient and low-cost embedded solution capable of converting text input into audible speech. The system is built around the ESP32 developed by Espressif Systems, which offers high processing capability along with built-in Wi-Fi and Bluetooth connectivity.

1.1 System Performance

The ESP32 microcontroller processes text data and converts it into speech signals. Its dual-core processor and sufficient memory allow the system to perform text processing and speech synthesis efficiently. The built-in communication modules enable wireless text input from external devices such as smartphones or computers.

1.2 Hardware Efficiency

The hardware design uses minimal components including the ESP32, an audio amplifier, and a speaker. This makes the system compact, energy-efficient, and cost-effective. The design also ensures easy integration with other IoT devices.

1.3 Reliability and Scalability

The system is designed to operate reliably for real-time speech generation. Because the ESP32 supports wireless connectivity, the system can be expanded to support cloud-based services, mobile applications, or smart home devices.

1.4 Power Consumption

The ESP32 is known for low power consumption, making the system suitable for portable or battery-powered applications such as assistive devices and smart notification systems.

2. Expected Outcomes

The proposed Text-to-Speech system is expected to achieve the following outcomes:

2.1 Successful Text-to-Speech Conversion

The system should successfully convert input text into understandable speech output through the connected speaker.

2.2 Real-Time Speech Generation

The system should generate speech with minimal delay after receiving text input, ensuring smooth real-time communication.

2.3 Clear and Audible Output

The audio output should be clear and audible with acceptable pronunciation accuracy.

2.4 Low-Cost Embedded Solution

The system should demonstrate that embedded platforms like the ESP32 can be used to build cost-effective speech systems.

2.5 Practical Applications

The system can be used in several real-world applications such as:
Assistive devices for visually impaired users Voice alert and notification systems
Smart home announcements educational tools

6. CONCLUSION

The Text-to-Speech (TTS) system using the ESP32 provides an effective and low-cost solution for converting written text into spoken audio. By utilizing the processing capabilities and wireless connectivity of the ESP32 developed by Espressif Systems, the system can efficiently process text input and generate understandable speech output through a speaker.The proposed system demonstrates how embedded platforms can be used to implement speech synthesis in compact and energy-efficient devices. The architecture allows real-time text processing and speech generation, making the system suitable for various applications such as assistive technologies, smart home notifications, and automated information systems.
Overall, the project proves that the ESP32 microcontroller is capable of supporting Text-to-Speech functionality in embedded systems. With its low power consumption, wireless communication features, and flexible programming environment, it provides a reliable platform for developing intelligent voice-based applications.

7. ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who helped me in completing my project titled **"Text to Speech System."** First of all, I would like to thank my project guide for providing valuable guidan ce, encouragement, and support throughout the development of this project. Their suggestions and advice helped me to understand the concepts and successfully complete the work. I would also like to thank our department and faculty members for providing the necessary facilities and resources required for this project. Their continuous support and motivation played an important role in the completion of this work. Finally, I would like to thank my friends and family members for their encouragement, cooperation, and moral support during the development of this project.

8. REFERENCES

[1] A. Sharma, R. Singh, P. Kumar, Prof. M. Deshmukh, Design and Implementation of Text to Speech System using ESP32 Microcontroller, International Journal of Engineering Research and Technology (IJERT), ISSN: 2278-0181, Volume 10, Issue 5, May 2021.

[2] S. Patel, N. Shah, K. Mehta, Prof. R. Tiwari, Development of Embedded Voice Alert System using ESP32, International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET), ISSN: 2319-8753, Volume 9, Issue 8, August 2020.

[3] M. Gupta, R. Verma, A. Jain, Prof. P. Kulkarni, Implementation of IoT Based Text to Speech Conversion System using ESP32, International Journal of Advanced Research in Computer Science (IJARCS), ISSN: 0976-5697, Volume 11, Issue 3, 2020.

[4] K. Reddy, S. Kumar, V. Prasad, Prof. D. Sharma, Embedded Speech Synthesis System using ESP32 for Assistive Applications, International Journal of Scientific and Technology Research (IJSTR), ISSN: 2277-8616, Volume 9, Issue 7, July 2020.

[5] P. Nair, A. Joseph, S. Thomas, Prof. R. Menon, Design of Low-Cost Text to Speech System using ESP32 for Smart Devices, International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249-8958, Volume 10, Issue 1, October 2020.