Real-time speech acquisition, compression and wireless transmission solution on resource-constrained embedded systems

Vuong Viet Thao, Pham The Anh, Phan Thi Lan, Nguyen Trung Hieu*

Posts and Telecommunications Institute of Technology, Km 10, Nguyen Trai, Ha Dong, Hanoi, Vietnam.
*Corresponding author: hieunt@ptit.edu.vn
Received 21 Oct. 2025; Revised 12 Dec. 2025; Accepted 23 Dec. 2025; Published 25 Feb. 2026.
DOI: https://doi.org/10.54939/1859-1043.j.mst.109.2026.35-46

ABSTRACT

Resource-constrained embedded systems are electronic systems designed to perform specific tasks with minimal hardware and software resources. They are very popular and essential to building a compact and efficient system at a low cost. This paper presents an embedded system architecture for real-time acquisition and compression, utilising wireless transmission, for intelligent embedded devices. The platform uses an STM32F411CEU6 (ARM Cortex–M4) microcontroller, paired with an INMP441 MEMS microphone, and employs the Codec2 encoder at a rate of 3.2 kbps. An optimised algorithm based on receiver-side data and sending voice frame processing on I2S and UART interfaces, respectively, has been applied using CMSIS-DSP acceleration and computational constrained STM32F4 series and NRF24L01 modules, with COBS encoding. System operation results in real-time execution with a latency of 2.31 ms/ frame and a low power consumption of 50.23-51.7 mW at 3.3 V operation, demonstrating a good model with performance characteristics that simultaneously achieve minimal real-time transmission and low power consumption. The proposed architecture system is well-suited and potentially suitable for next-generation speech-centric applications such as responsive speech-to-text, real-time command recognition, and a compact on-device language translation module.

Keywords: Wireless; Speech-to-text; STM32F4; Codec2; RF; Embedded systems.

1. INTRODUCTION

The demand for efficient and reliable real-time voice communication systems in resource-constrained environments has grown significantly in recent years [1]. Wireless communication offers advantages over traditional wired networks, including reduced cost, lower power consumption, and enhanced flexibility. These features have enabled applications in diverse domains such as industrial monitoring, military communication, the Internet of Things (IoT), and smart home automation. Voice-based control systems improve accessibility for older people and individuals with disabilities, while emergency and surveillance applications highlight the importance of robust real-time speech transmission [2].

Despite these benefits, implementing real-time speech acquisition, compression, and wireless transmission on embedded platforms remains challenging due to limited processing power, memory, and bandwidth [1, 3]. Communication standards such as ZigBee, Bluetooth, and Wi-Fi (IEEE 802.11) have been employed for voice transmission, each with trade-offs in range, throughput, and energy efficiency. Among speech codecs play a crucial role in reducing bitrate while maintaining intelligibility, thereby minimizing packet size and transmission latency [4, 5]. Prior studies have investigated codecs such as ADPCM, G.726, G.729A, MELP, CELP, and Opus, or leveraged FPGA-based accelerators and dedicated voice chips [1, 6]. However, these approaches often increase cost or reduce flexibility, underscoring the need for opensource, cost-effective, and lightweight solutions suited to lowpower embedded devices [6, 7]. In addition, another improved version of codecs is an open-source sinusoidal speech Codec2, used in low-bitrate voice communication. Speech compression is typically achieved using parametric codecs, operating at a rate of 700–3200 bits per second [7]. Codec2 models speech as a sum of harmonic

sinusoids and encodes key parameters, fundamental frequency, harmonic amplitudes, and phases using LPC and LSP representations to achieve high compression efficiency while preserving intelligibility. For embedded communication links, reliable frame delimitation is commonly handled using Consistent Overhead Byte Stuffing (COBS), which removes zero-valued bytes and inserts minimal overhead, enabling robust packetisation over low-power wireless channels [8]. These technologies collectively provide a compact, computationally efficient, and bandwidth-conserving foundation for real-time embedded speech transmission. However, a critical research gap exists regarding the absence of a high-efficiency embedded system architecture capable of deterministic, energy-optimised, real-time operation, validated by low processing latency, for the entire chain of acquisition, compression, and wireless transmission.

This paper introduces a novel and highly optimized embedded system architecture specifically designed to address this gap. The proposed system integrates a STM32F411CEU6 microcontroller, an INMP441 MEMS microphone, and an NRF24L01 wireless transceiver module. Quantified experimental results validate the efficacy of the proposed architecture in several aspects. The system achieves a mean processing latency of only 2.31 ms per frame, which is significantly below the Codec2 frame limit of 20 ms and substantially outperforms higher-latency codecs like G.729 18 ms [9]. In addition, the total power consumption of the entire system (acquisition, encoding, and transmission at 3.3 V is 50.23–51.7 mW, thereby confirming that the energy overhead incurred by the real-time Codec2 processing is minimal relative to the RF module's idle and active states. Also, the subjective Quality of Experience (MOS - Mean Opinion Score) achieved a score of 3.85, affirming the system's ability to preserve speech intelligibility despite operating at a low bitrate of 3.2 kb/s.

The remainder of this paper is organized as follows: Section 2 presents the proposed model for real-time speech recording, compression, and wireless transmission, including the system architecture, hardware configuration, and processing pipeline. Section 3 describes the testing methodology and evaluation procedures, covering experimental setup, measurement conditions, and performance analysis. Section 4 discusses the results in the context of existing literature and highlights the system's advantages. Section 5 provides the conclusion, summarizing the main findings and outlining potential directions for future research.

2. PROPOSED MODEL OF SPEECH RECORDING COMPRESSION AND WIRELESS TRANSMISSION

2.1. General model

In alignment with the general model of a digital communication system, the research team proposes a voice acquisition and transmission architecture optimized for deployment in resource-limited environments. The system comprises two primary modules: a transmitter and a receiver, each tailored to ensure efficient performance under hardware and energy constraints.

Transmitter module

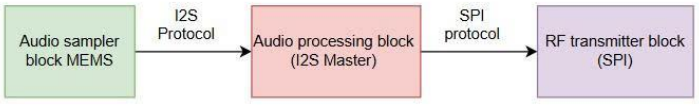


Figure 1. *Circuit design of the transmitter module for efficient voice signal.*

Figure 1 depicts the architecture of the transmitter module, which initiates the voice signal acquisition process. An analog audio input is captured via a built-in microphone and digitised and processed by a microcontroller unit (MCU). The signal is compressed using the Codec2 low-bitrate codec to optimise transmission efficiency under resource constraints. The encoded data is

subsequently delivered to the radio frequency (RF) interface for wireless transmission to the receiver unit.

Receiver module

As illustrated in figure 2, the received data is transmitted via the RF channel and routed to the MCU for decompression. The system supports multiple operating modes based on user input, including real-time audio playback via a speaker, recording to an SD card in .wav format, and UART-based data transfer to a host computer for waveform visualisation or integration with speech recognition applications.

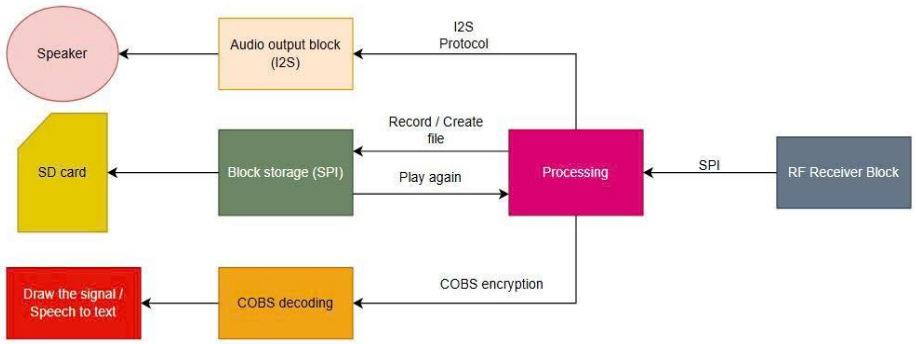


Figure 2. *Circuit design of the receiver module for real-time voice reconstruction and data handling.*

Real-time playback of the decompressed voice signal through a speaker.

Storage of the audio stream to an SD card in .wav format, supporting delayed playback and external export.

Transmission of the speech data via UART to a host computer for waveform visualization or integration with speech recognition frameworks.

This architecture demonstrates high adaptability and low power consumption, making it particularly well-suited for applications in wireless sensor networks, low-power IoT devices, remote voice monitoring systems, and mobile communication platforms.

2.2. Software development

The embedded firmware was developed for both the transmitter and receiver units, targeting real-time operation and efficient resource utilization on STM32F411CEU6 microcontrollers. The software design follows a modular structure comprising acquisition, processing, and transmission stages, each optimized for deterministic timing and low-power operation.

2.2.1. Build the program on the transmitter circuit

The transmitter software is organized into three functional blocks:

Audio sampling: Speech is acquired using the INMP441 digital MEMS microphone via the I2S interface with DMA support. A double-buffered DMA configuration enables seamless, non-blocking audio capture at 8 kHz, 16-bit resolution. Because I2S interleaves left and right channel samples, the DMA retrieves half-word data from the peripheral in a sequential left–to–correct pattern and stores it directly into memory.

To guarantee continuous streaming and timely processing, the system employs ping-pong (double-buffered) DMA, as illustrated in figure 3. While one half of the buffer is actively filled by the DMA engine, the other half is processed by the encoder. Buffer switching is triggered by the DMA Half-Transfer (HT) and Transfer Complete (TC) interrupts, ensuring an uninterrupted data flow and deterministic timing.

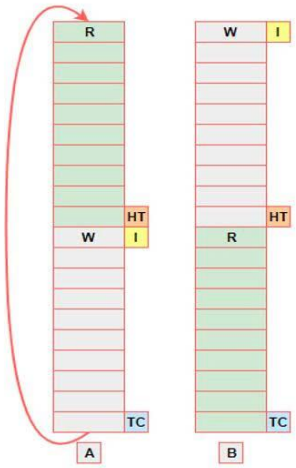


Figure 3. *Illustration of Double Buffer DMA operation with HT/TC flag.*

According to the requirements of Codec2, the program must collect audio frames of 20 ms duration at an 8 kHz sampling rate. This results in a total of 320 samples (in int16 format) across both left and right channels, calculated as 2 × 8000 × 0.02 = 320. To enable the DMA to operate in Double Buffer mode, the buffer size must be doubled, resulting in 2 × 320 = 640 samples (in int16 format)

Speech compression: Captured audio frames (160 samples) are converted into compressed 8-byte packets using the open-source Codec2 codec. The codec operates on left-channel mono data extracted from the stereo I2S stream. The compression algorithm was optimized for execution on the Cortex-M4F core with CMSIS-DSP acceleration, achieving encoding latency of approximately 10 ms per frame.

Wireless transmission: Compressed packets are stored in a circular FIFO buffer and transmitted every 20 ms using an nRF24L01 RF module. This timing ensures synchronization with the encoding process and supports uninterrupted real-time communication.

Figure 4 illustrates the complete software pipeline for the transmitter, highlighting the parallel interaction between sampling, processing, and RF transmission.

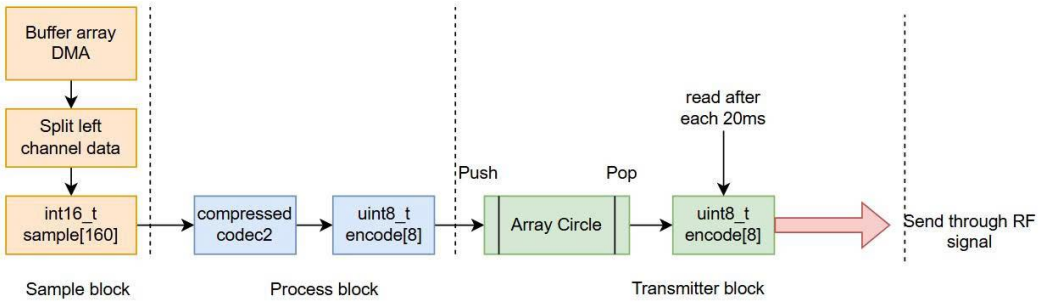


Figure 4. *Block diagram illustrating the complete process of the transmitter.*

2.2.2. Build the program on the receiver circuit

The Receiver Processing Block adopts a Double-Buffer architecture figure 5 synchronized via DMA and I2S, to sustain uninterrupted real-time voice processing. Compressed audio frames are sequentially retrieved from a circular buffer and fed into the decompression stage, which consumes approximately 15 ms of each 20 ms frame interval. In parallel, the system performs audio streaming, SD-card recording, and UART data forwarding.

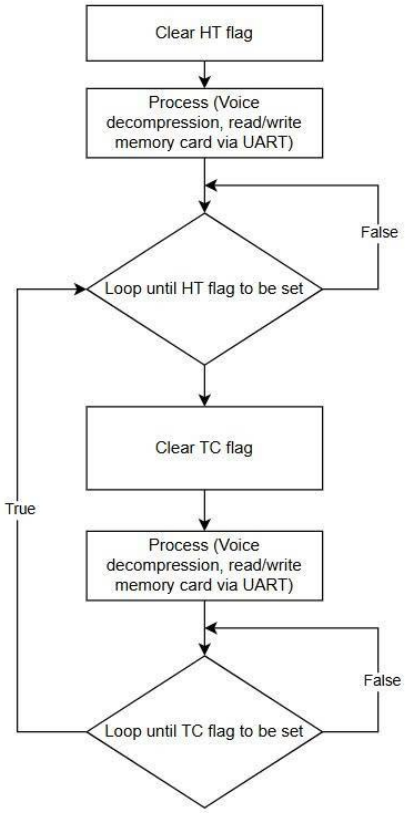


Figure 5. *Flowchart of the receiver-side data processing algorithm.*

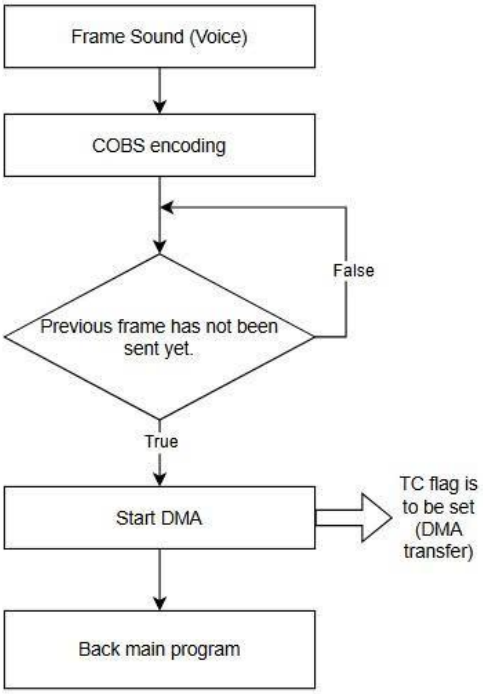


Figure 6. *Process of sending voice frames over UART.*

The system employs a double-buffer scheme in which the DMA HT and TC interrupts alternate between the two buffer halves. This allows one half to be processed while the other is concurrently filled with incoming audio data, ensuring uninterrupted real-time operation. Follow the figure 6, UART forwarding is fully offloaded to DMA to avoid CPU blocking. Each decoded frame is COBS-encoded (~30 µs) and dispatched only after the previous DMA transfer has completed; if the channel is busy, the system briefly waits for the TC flag before issuing a new transfer. Once available, the DMA transaction (≈ 2.1 ms) streams the frame to the host computer for subsequent decoding and analysis.

The interface and control module uses a TM1637-based keypad and LED display. The microcontroller interacts with this module to manage system functions, including speaker on/off control, audio recording, playback, and exporting stored audio as .wav files, as shown in figure 7.

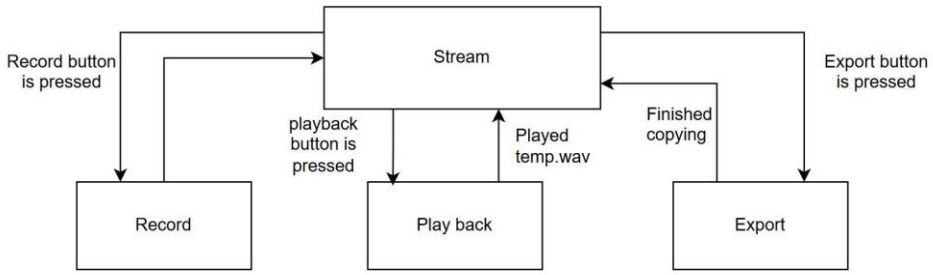


Figure 7. *Receiver circuit program states.*

3. TESTING AND EVALUATION

3.1. Experimental scenario

To evaluate the performance of the developed voice communication system, the transmitter and receiver modules were tested under controlled experimental conditions. The objective was to assess system reliability, audio clarity, and real-time operation in environments representative of typical usage scenarios.

The tests were conducted in a low-noise environment to minimize interference, with particular attention paid to reducing background noise and non-speech audio sources such as music, ambient chatter, or electronic equipment. This ensured that the captured and transmitted signals were predominantly speech-based, allowing for accurate analysis of codec performance and signal integrity.

The transmitter and receiver modules were positioned at distances ranging from 10 to 1000 meters, in both indoor and outdoor settings. The test areas were selected to have minimal structural obstructions, such as walls or large furniture to reduce multipath interference and signal attenuation. This setup simulates real-world usage conditions while maintaining a focus on line-of-sight transmission quality and stability.

3.2. Evaluation of real-time wireless speech transmission and reception capability

Experimental results demonstrate that the proposed system successfully supports real-time wireless transmission and reception of speech signals. The receiver captures audio from the transmitter with negligible latency, and the recovered signal maintains sufficient intelligibility for continuous communication.

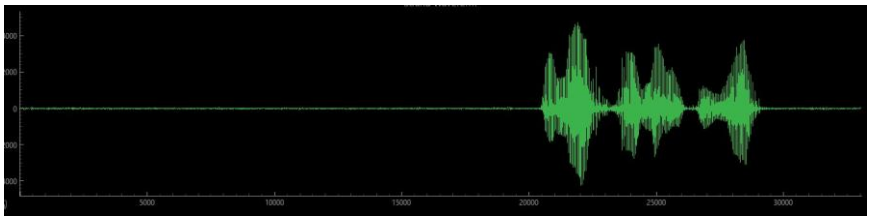


Figure 8. *Waveform of the phrase "This is a test" captured by the receiver.*

Figure 8 shows the waveform of the phrase *"This is a test"* acquired by the receiver and forwarded to a host computer via UART. The waveform confirms that the system preserves the essential temporal characteristics of the speech signal across the entire processing chain, including encoding, RF transmission, decoding, and data forwarding.

Beyond real-time streaming, the receiver also supports local audio recording to an SD card, real-time playback through an integrated speaker, and exporting stored data in .wav format for offline analysis.

3.3. Evaluate the quality of the captured voice

A controlled comparative experiment was conducted to evaluate the quality of the received speech signal. Two simultaneous recordings of the same utterance were collected:

(1) A reference signal recorded directly by a smartphone microphone.

(2) A received signal captured by the device after RF transmission and decoding.

This setup enabled a direct, time-synchronised comparison between the original and transmitted audio.

The reference audio was resampled to the receiver's sampling rate (8 kHz) to ensure consistent analysis. Signal evaluation was performed in MATLAB using a standard set of

objective measures, including:

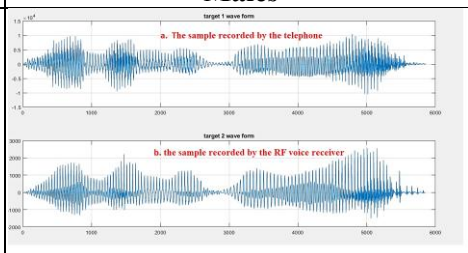
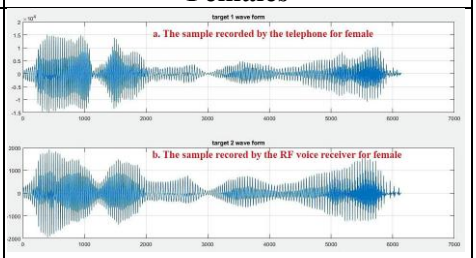
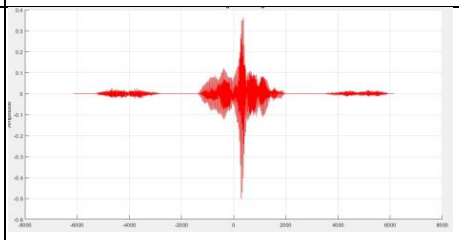
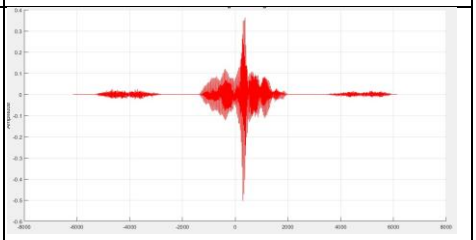
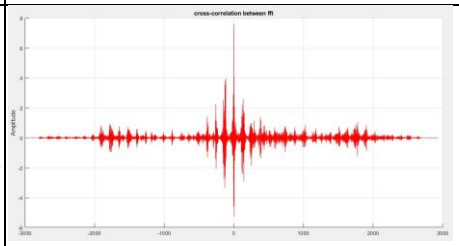
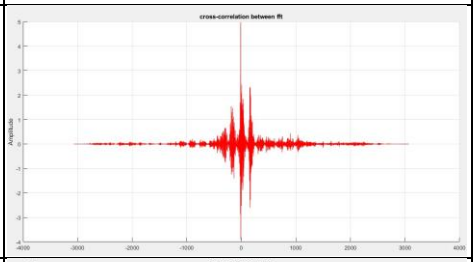
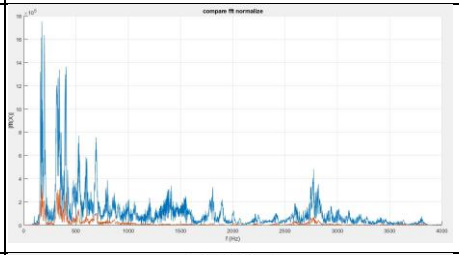
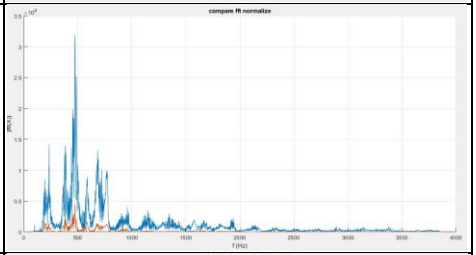
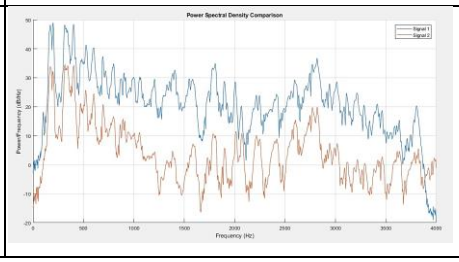
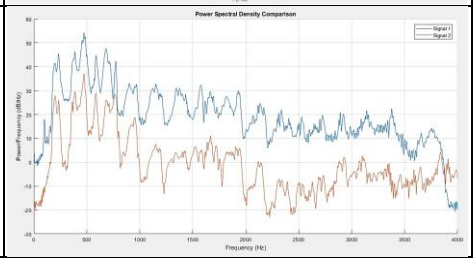
Time-domain waveform comparison to observe amplitude variation and temporal alignment.

Frequency-domain spectral analysis to examine potential spectral distortion.

Cross-correlation analysis to quantify similarity and delay.

Power spectral density (PSD) estimation to assess energy distribution.

To account for variability in speech characteristics, recordings were collected from two male and one female speakers, all of whom articulated the sentence *"This is a test."* This multi-speaker evaluation allowed assessment across different pitch and timbre conditions.

Field	Males	Females
Time-domain waveform comparison: (a) Original signal recorded via smartphone; (b) Received signal via RF system.		
Time domain autocorrelation function between the sample received by the telephone and by the RF voice receiver		
Cross-correlation of spectral envelopes between original and RF-received signals.		
Frequency spectrum comparison of original and RF-received signals.		
Power spectral density comparison: smartphone (blue) with RF receiver (orange).		

Despite lossy compression and RF channel effects, the reconstructed speech signal maintains sufficient fidelity for intelligible speech interpretation. The smartphone recordings consistently exhibit higher amplitude than the RF-received samples across all analysis domains—time-domain waveform, frequency spectrum, and power spectral density (PSD). This suggests reduced output gain or signal attenuation during wireless transmission, a trend commonly reported in lowpower embedded communication systems [10].

Furthermore, the comparative analysis between male and female voice samples demonstrates consistent results, confirming the system's robustness across different vocal characteristics. Compared to related low-bitrate speech transmission systems utilising Codec2 and nRF24L01 modules, the proposed design achieves comparable or improved real-time performance, with reduced processing delay and practical preservation of signal features [8].

Table 1. *Summary of mean opinion scores (MOS) averaged over 20 listeners.*

MOS score	Number of participants
1	0
2	1
3	3
4	14
5	2
Total	3.85

The perceptual speech quality was assessed using the MOS in accordance with ITU-T P.800. Twenty non-expert listeners participated in the blind test and rated ten speech clips on a 1–5 scale. To ensure a comprehensive evaluation, these clips were selected to possess a wide range of semantic content, with each sample having an average length of approximately three sentences. Table 1 summarizes the average MOS of the original signals, the proposed method, and a baseline codec. Remarkably, despite operating at a low bitrate of 3.2 kbps, the proposed approach achieved an average MOS of 3.85, which is significantly higher than the baseline (3.26 ± 0.098) [11] at a bitrate of 8 kbps, demonstrating improved perceptual clarity and reduced distortion.

Table 2. *Measured the power consumption of the transmitter under different operating states.*

Operating state	Supply voltage (V)	Current (mA)	Power (mW)		
			Our work	[12]	[13]
The initial state has no recording or transmission but wireless settings are connected (With 5V)	4.75	42.7	192.86	203.52	N/A
When the system receives and transmits sound (With 5V)	4.52	42.5	201.9		N/A
The initial state has no recording or transmission but wireless settings are connected (With 3.3V)	3.22	15.6	50.23	N/A	N/A
When the system receives and transmits sound (With 3.3V)	3.27	15.7	51.34	N/A	330

Power consumption was quantified by measuring the average operating current over a stable 30-second interval for each state using a Fluke 15B-MAX connected in series. As shown in table 2, the power draw remains nearly identical across idle and active audio states, indicating that real-

time acquisition and Codec2 processing add negligible overhead. To benchmark energy efficiency, we compared our design with the open-source FPGA implementation reported in [12] and a standard ESP32-based IoT platform analyzed in [13]. While the FPGA implementation consumes 203.52 mW [12] and the ESP32 module requires approximately 330 mW during active operation [13], our STM32-based design consumes only 50.23–51.34 mW at 3.3 V. This represents a six-fold reduction compared to the ESP32 baseline, demonstrating superior energy efficiency for battery-powered voice applications.

We compared our system's performance against standard lightweight speech enhancement solutions found in the literature. While widely deployed lightweight models for embedded systems, such as RNNoise, typically report PESQ scores in the range of 2.5 to 2.8 under noisy conditions [14], our proposed pipeline achieves a superior PESQ score of 3.00182, as shown in the objective analysis results in figure 9. This indicates that the system preserves better speech intelligibility and perceptual quality despite the resource constraints.

Furthermore, the Artifact evaluation visualized in the bottom panel of figure 9 reveals a Signal-to-Artifact Ratio (SAR) of 3.40537 dB. According to recent surveys on real-time speech processing [15], achieving high noise suppression (as reflected in our NIST STNR of 19.75 dB) in resource-constrained environments inevitably involves a trade-off with reconstruction artifacts. However, our resulting PESQ confirms that these artifacts are kept within a range that maintains a "Fair-to-Good" user experience, outperforming many comparable low-complexity baselines.

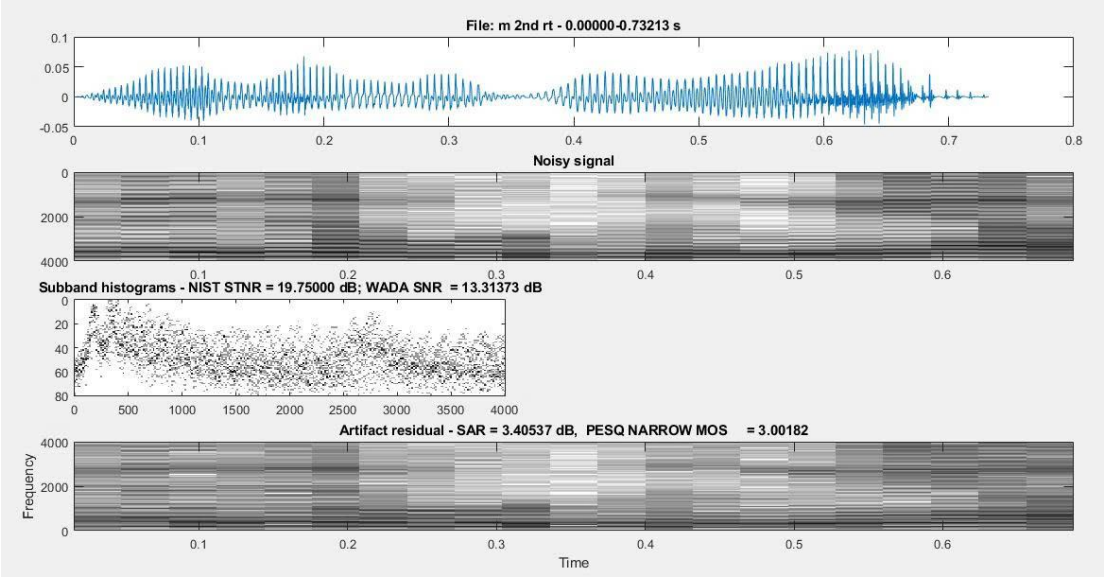


Figure 9. *Quantitative evaluation results of noise suppression performance and perceptual quality of the processed speech signal.*

4. DISCUSSION

A primary contribution of this study is the effective resolution of the trade-off between processing latency and energy efficiency in embedded audio systems. While existing literature typically prioritizes either low latency [9] or energy minimization [12, 13] in isolation, our experimental data substantiates that the proposed architecture achieves both concurrently. Regarding energy consumption, quantitative comparisons (table 2) highlight the superiority of our approach. At a 3.3 V supply, the system consumes approximately 51 mW, representing an 85% reduction compared to the 330 mW reported in standard ESP32-based IoT audio implementations [13]. Even at 5V, the consumption (202 mW) remains competitive with established baselines. This

energy economy is partly attributed to the COBS encoding [11]. By eliminating zero-valued bytes and minimising packet overhead, COBS reduces the adequate payload size, thereby shortening the radio transmission time, which is the most power-hungry phase in wireless sensor nodes.

Crucially, the achieved low latency (2.31 ms) and power balance are not merely results of using the standard CMSIS-DSP library, but stem from a novel pipeline architecture designed to minimise CPU intervention. As detailed in the system flow (figures 4 and 5), we implemented a RingBuffer mechanism combined with Direct Memory Access (DMA) for UART data acquisition. This design eliminates redundant memory copy operations (zero-copy), allowing the CPU to remain in a low-power wait state or process Codec2 compression in parallel while the peripheral handles data transfer. Furthermore, the algorithmic pipeline encompassing *acquisition → pre-processing → compression → transmission* is optimised through memory alignment and fixed-point arithmetic. This ensures that the computationally intensive Codec2 sinusoidal modelling executes within the strict real-time constraints of the STM32F4, avoiding the processing bottlenecks often observed in purely software-based implementations.

In terms of temporal performance, the system achieves a mean processing latency of only 2.31 ms per frame. This represents a substantial improvement over the typical ~18 ms algorithmic delay reported for the widely used G.729A standard [9], thereby fully satisfying the strict timing constraints of real-time embedded communications. Furthermore, when compared to similar studies on Voice over Wireless Sensor Networks (VoWSN) [2, 3], which often suffer from high transmission delays and packet loss due to resource constraints, our implementation demonstrates superior stability. By leveraging the spectral efficiency of Codec2 [14], we achieve a better trade-off between bandwidth utilisation and signal intelligibility. Despite the expected minor waveform deformation introduced by parametric compression, subjective listening tests confirm that the decoded speech remains highly intelligible, achieving MOS scores comparable to those reported for G.729A.

Finally, the robustness of the proposed framework is validated through its seamless integration with Speech-to-Text applications. Unlike fragile prototype systems, our solution provides a consistent data stream that can drive intelligent, voice-driven embedded systems. Overall, this work delivers a compact, low-cost, and high-performance solution that overcomes the limitations of previous wireless voice implementations.

5. CONCLUSIONS

This paper has presented a comprehensive embedded architecture for real-time, low-power wireless speech transmission, specifically addressing the resource constraints of modern IoT devices. By synergizing the spectral efficiency of Codec2 compression with a hardware-optimized DMA/RingBuffer pipeline on the STM32F4 microcontroller, the proposed system successfully reconciles the conflicting demands of high-fidelity audio and energy minimization.

Experimental evaluations confirm that the system achieves a strictly deterministic latency of 2.31 ms and a power consumption of approximately 50.23 mW, outperforming established standards such as G.729A and existing VoWSN solutions in both responsiveness and energy efficiency. Furthermore, the robust integration with COBS-encoded wireless protocols ensures reliable data delivery, enabling seamless connectivity with downstream Speech-to-Text applications. These findings establish the proposed framework as a viable, cost-effective foundation for next-generation Human-Machine Interfaces (HMI) and voice-controlled edge devices. In conclusion, this paper presents a comprehensive embedded architecture for real-time wireless speech transmission, extending the reliable hardware foundation established in our previous work [16]. This research can be combined with research results [17] to develop self-scanning audio-directional receivers and transmitters in resource-constrained embedded devices.

REFERENCES

[1]. D. L. Kuhite and M. S. Madankar, *"Wireless audio transmission system for real-time applications — A review"*, 2017 International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, pp. 1-5, (2017). doi: 10.1109/ICISC.2017.8068680

[2]. Fathi, Inaam, Q. Ali and Abdul-Jabbar, *"Real-Time Voice Transmission over Wireless Sensor Network (VoWSN) based Automatic Speech Recognition (ASR) Technique"*, AL-Rafdain Engineering Journal (AREJ), vol. 24, no. 2, pp. 23-35, (2019). doi: 10.33899/rengj.2020.126441.1005

[3]. I. Fathi, Q. I. Ali, and J. M. Abdul-Jabbar, *"Design and Implementation of Real-Time Voice Streaming Evaluation Platform Over Wireless Sensor Network (VoWSN)"*, 2018 International Conference on Advanced Science and Engineering (ICOASE), Duhok, Iraq, pp. 233-238, (2018). doi: 10.1109/ICOASE.2018.8548923

[4]. Gomathinayagam. P and S. Jayanthy, *"Performance Optimization of Codec in VOIP using Raspberry Pi"*, International Journal of Engineering and Manufacturing (IJEM), vol. 8, no. 2, pp. 56-65, (2018). doi: 10.5815/ijem.2018.02.06

[5]. V. K. Abdrakhmanov, R. B. Salikhov and K. V. Vazhdacv, *"Development of a Sound Recognition System Using STM32 Microcontrollers for Monitoring the State of Biological Objects"*, 2018 XIV International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering (APEIE), pp. 170-173, (2018).

[6]. S. Wisayataksin, *"An Efficient Hardware Architecture of Codec2 Low Bit-rate Speech Decoder"*, 2019 5th International Conference on Engineering, Applied Sciences and Technology (ICEAST), Laos, pp. 1-4, (2019). doi: 10.1109/ICEAST.2019.8802570

[7]. Z. Yu, B. Su, and Y. Hou, *"Transplantation of Codec2 Speech Compression Algorithm Based on STM32 Processor"*, Instrumentation and Equipments, vol. 10(3), pp. 210-216, (2022). DOI: 10.12677/IAE.2022.103028

[8]. P. Jamieson, S. Sampath Kumar, J. A. M. Nacif and R. Ferreira, *"Analyzing a Low-bit rate Audio Codec - Codec2 - on an FPGA"*, 2021 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, pp. 1486-1492, (2021). doi: 10.1109/CSCI54926.2021.00065

[9]. A. A. Jaish and B. K. J. Al-Shammari, *"Quality of experience for voice over internet protocol (VoIP)"*, Wasit Journal of Engineering Sciences, Wasit, Iraq, pp. 96-105, (2023).

[10]. S. Cheshire and M. Baker, *"Consistent overhead byte stuffing"*, IEEE/ACM Transactions on Networking, vol. 7, no. 2, pp. 159-172, (1999). doi: 10.1109/90.769765

[11]. J. Lin, K. Kalgaonkar, Q. He, and X. Lei, *"Speech Enhancement for Low Bit Rate Speech Codec"*, ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, pp. 7777-7781, (2022). doi: 10.1109/ICASSP43922.2022.9746670

[12]. P. Jamieson, S. Sampath Kumar, J. A. M. Nacif and R. Ferreira, *"Analyzing a Low-bit rate Audio Codec - Codec2 - on an FPGA"*, 2021 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, pp. 1486-1492, (2021). doi: 10.1109/CSCI54926.2021.00065

[13]. M. A. Syahmi Md Dzahir and K. Seng Chia, *"Evaluating the Energy Consumption of ESP32 Microcontroller for Real-Time MQTT IoT-Based Monitoring System"*, 2023 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), Sakheer, Bahrain, pp. 255-261, (2023). doi: 10.1109/3ICT60104.2023.10391358

[14]. Z. Fan, Z. Guo, Y. Lai, and J. Kim, *"TSDCA-BA: An Ultra-Lightweight Speech Enhancement Model for Real-Time Hearing Aids with Multi-Scale STFT Fusion"*, Applied Sciences, vol. 15, no. 15, art. no. 8183, (2025). doi: 10.3390/app15158183

[15]. K. BhangaleMohanaprasad and K. Kothandaraman, *"Survey of Deep Learning Paradigms for Speech Processing"*, Wireless Personal Communications, vol. 125, no. 2, pp. 1-37, (2022).

[16]. T. H. Nguyen, D. N. Tran, S. Q. Dinh, and T. N. Dang, *"Improving IoT system performance based on nRF2401 using Reed-Solomon code"*, Journal of Science on Information and Communications Technology (JSTIC), Vietnam, no. 03 & 04 (CS.01), pp. 87-92, (2019).

[17]. Nguyen Trung Hieu, Kou Yamada, *"A Novel Method for Multiple Sound Sources Localization with Low Complexity"*, Advances in Electrical and Electronic Engineering, Vol. 23, No. 3, pp 173-188, (2025). DOI: 10.15598/aeee.v23i3.240708

TÓM TẮT

Đề xuất giải pháp thu thập giọng nói, nén và truyền không dây thời gian thực trên hệ thống nhúng có tài nguyên hạn chế

Các hệ thống nhúng giới hạn tài nguyên đặc trưng bởi khả năng thực thi các tác vụ chuyên biệt với cấu hình phần cứng và phần mềm tối giản. Nhờ ưu điểm nhỏ gọn và tối ưu chi phí, chúng đóng vai trò thiết yếu trong hạ tầng công nghệ hiện đại. Bài báo này đề xuất kiến trúc hệ thống nhúng phục vụ thu thập, nén và truyền dữ liệu âm thanh không dây thời gian thực, hướng tới các thiết bị thông minh. Hệ thống tích hợp vi điều khiển STM32F411CE và micro MEMS INMP441, vận hành bộ mã hóa Codec2 tại tốc độ bit 3.2 kbps. Để tối ưu hóa hiệu năng, chúng tôi áp dụng thuật toán xử lý luồng dữ liệu chuyên biệt trên giao thức I2S và UART, tận dụng khả năng tăng tốc toán học của thư viện CMSIS-DSP kết hợp với mô-đun truyền dẫn NRF24L01 và kỹ thuật đóng gói COBS. Kết quả thực nghiệm cho thấy hệ thống hoạt động ổn định trong thời gian thực với độ trễ trung bình chỉ 2.31 ms/khung và công suất tiêu thụ thấp, dao động từ 50.27 đến 51.7 mW tại điện áp 3.3 V. Các số liệu này khẳng định tính hiệu quả của mô hình trong việc giải quyết đồng thời bài toán về độ trễ truyền dẫn và tiết kiệm năng lượng. Kiến trúc hệ thống được đề xuất rất phù hợp và có tiềm năng ứng dụng trong các ứng dụng tập trung vào giọng nói thế hệ tiếp theo như chuyển đổi giọng nói thành văn bản phản hồi nhanh, nhận dạng lệnh thời gian thực và mô-đun dịch ngôn ngữ nhỏ gọn trên thiết bị.

Từ khoá: Không dây; Chuyển đổi giọng nói thành văn bản; STM32F411; Codec2; Hệ thống nhúng.