World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

Check for updates

Design and implementation of an AI-based wireless real-time voice translation system with directional audio output

Sundaresh Jeyaram *, Akila Deshan, Siva Uthayaraj and Rajan Vinojan

Department of Physical Science, Faculty of Applied Science, Trincomalee Campus, Eastern University, Sri Lanka, Nilaveli, 31010, Trincomalee, Sri Lanka.

World Journal of Advanced Engineering Technology and Sciences, 2025, 16(03), 032-037

Publication history: Received on 26 July 2025; revised on 31 August; accepted on 03 September 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.16.3.1324

Abstract

This research presents the design and implementation of an AI-driven wireless real-time voice translation system incorporating directional acoustic output, designed to facilitate seamless multilingual communication in dynamic environments. The proposed architecture integrates real-time speech recognition, Neural Machine Translation (NMT), and spatially controlled audio synthesis within a unified framework. Voice input is captured via a Frequency Modulated (FM) wireless microphone and transmitted to a Python-based desktop platform. The signal undergoes Automatic Speech Recognition (ASR) using a deep learning-based Speech-To-Text (STT) engine, followed by semantic translation via Google's NMT-API, leveraging transformer-based models for high contextual fidelity.

The translated linguistic output is rendered into naturalistic human like speech through a neural Text-To-Speech (TTS) engine and delivered via a parametric speaker array utilising ultrasonic transducers with 40 kHz Pulse Width Modulation (PWM). This enables highly directional audio propagation with minimal ambient leakage, ensuring privacy and intelligibility for the intended listener without the need for wearable audio devices. A resource constrained ESP32 microcontroller orchestrates real-time data acquisition, translation synchronisation, and modulation control for the parametric output. Empirical evaluation demonstrates low end-to-end latency (1.5 - 2.5 seconds) and high ASR accuracy (90–95%), validating the system's viability for deployment in multilingual conferences, educational domains, and public communication interfaces.

Keywords: Real-time Speech Translation; Neural Machine Translation (NMT); Directional Audio; Speech-To-Text (STT); Text-To-Speech (TTS); Acoustic Beamforming

1. Introduction

In an era of unprecedented global connectivity, the ability to overcome linguistic barriers is paramount to facilitating effective communication, particularly in multilingual settings such as international conferences, academic institutions, public information systems, etc. Conventional translation solutions often involve manual intervention or lack real-time capability, and need of external wearable audio devices, thereby limiting their scalability and usability in dynamic, real-world scenarios.

This paper introduces the design and implementation of an AI-enhanced, wireless, real-time speech translation system with directional audio output, developed to address these limitations. The proposed system integrates advanced modules for speech acquisition, natural language processing, and spatially targeted audio delivery within a compact and cost efficient embedded platform. The architecture is divided into three primary sub-systems such as Input, Processing, and Output, as illustrated in the functional block diagram of the system in Figure 1.

* Corresponding author: Sundaresh Jeyaram

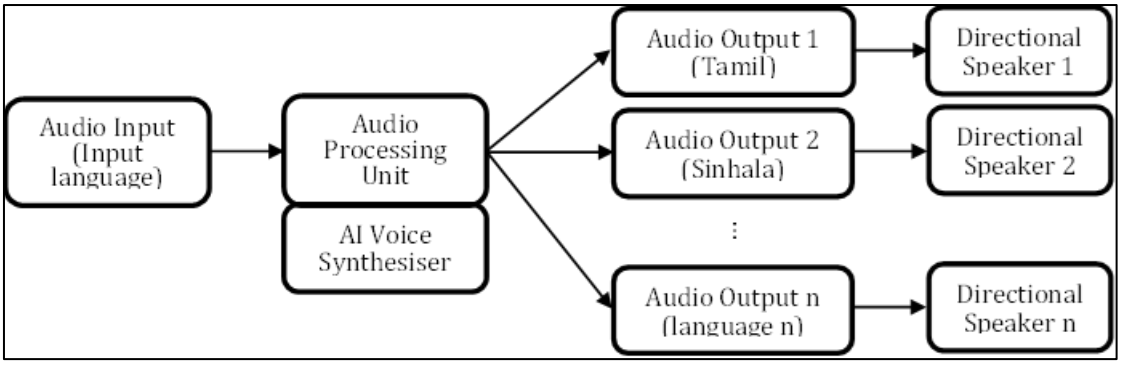

Figure 1 Block diagram of the system

In the block diagram, the languages indicated in parentheses represent those used for empirical validation and demonstration; however, the system architecture is inherently scalable and supports the full range of languages available via the Google Translate API, enabling global language coverage without modification to the core software.

The Input Module employs an FM based wireless microphone system for untethered speech acquisition, transmitting analog audio signals to a desktop receiver. The signal is subsequently processed by a Python-based application that orchestrates a pipeline consisting of Automatic Speech Recognition (ASR), NMT, and TTS synthesis. Deep learning models are utilised at each stage to ensure high accuracy and low latency.

The resulting translated speech is emitted through a custom made parametric speaker array, utilising ultrasonic transducers driven by precisely modulated 40 kHz PWM signals. The ESP32 microcontroller, operating as a real-time PWM generator and controller, orchestrates the ultrasonic carrier signal generation. This configuration enables acoustic beamforming, allowing highly focused sound projection to a specific spatial target. The directional nature of the parametric audio minimises cross-talk and ensures privacy and clarity for the intended listener, eliminating the need for conventional headsets or earphones in shared environments.

Objectives

The principal objectives of this research focus on the design and implementation of a low-latency, real-time multilingual speech translation pipeline by leveraging advanced AI techniques, including ASR, NMT, and neural TTS synthesis. To develop a wireless voice input mechanism utilising FM transmission to facilitate untethered operation and enhance system mobility across diverse environments. Additionally, the research explores the use of parametric loudspeaker technology, based on ultrasonic transducers, to deliver highly directional and non-intrusive acoustic output, thereby minimising auditory spillover and preserving listener privacy. An ESP32 microcontroller is integrated into the system to generate stable, high-frequency 40 kHz PWM signals required for precise ultrasonic carrier modulation and parametric sound synthesis. Finally, a compact, low-cost, and user-accessible prototype is constructed, aimed at practical deployment in real-world multilingual communication settings such as conferences, museums, classrooms, and public venues.

1.1. Real-time Multilingual Translation

The field of machine translation has evolved significantly, from early rule-based systems to modern NMT. Rule-based approaches [1] were limited by inflexibility and vocabulary constraints. Statistical Machine Translation (SMT) [2], which emerged in the 1990s, offered probabilistic models but struggled with context accuracy. NMT has since revolutionised the field with deep learning techniques that offer superior translation quality and context sensitivity [3]. STT technologies now integrate Natural Language Processing (NLP) with AI to deliver real-time, spoken multilingual communication [4].

1.2. Wireless Audio Systems

Wireless audio systems began with analog AM/FM technologies [5], designed mainly for broadcasting. The development of Bluetooth and Wi-Fi enabled higher fidelity and lower latency communication for personal and professional requirements [6]. Modern wireless microphones provide reliable voice input in dynamic environments, significantly improving user mobility [7].

1.3. Ultrasonic Parametric Speakers

Parametric speakers operate by utilising ultrasonic waves, typically in the range of 40 kHz to 60 kHz, to modulate high-frequency signals that are subsequently demodulated by the air to produce audible sound [8]. This approach enables highly directional sound projection, significantly reducing acoustic spillover and enhancing listener specificity. While ultrasonic waves were initially employed in sonar and military technologies, their application has since expanded to include parametric speaker systems in contemporary settings such as museums, retail environments, and personal audio devices [9].

1.4. Signal Processing and Microcontroller Integration

Signal processing is essential in audio systems for ensuring clarity, minimising noise, and enabling real-time performance [10]. Microcontrollers like the ESP32 are widely adopted for their efficiency, low power consumption, and versatility in signal modulation tasks. Specifically, the ESP32's higher resolution of Analog to Digital Conversion (ADC) and PWM generation capability is vital for driving ultrasonic transducers used in parametric speakers [11,12].

2. Materials and Methods

The system was developed using a structured, modular design methodology, beginning with the identification of functional and performance requirements. This was followed by the systematic selection of hardware and software components, the integration of signal processing and control subsystems, and iterative testing under real-world conditions. The overarching objective of the system is to facilitate real-time multilingual translation from wireless voice input, coupled with spatially targeted audio output through parametric acoustic technology.

2.1. System Architecture and Workflow

The architecture of the proposed system is organized into three interdependent functional modules: Input, Processing, and Output as illustrated in Figure 1. These modules collectively enable continuous, low-latency translation and delivery.

The Input Module comprises a wireless FM microphone that transmits analog voice signals to a corresponding FM receiver. This receiver is interfaced with a host computing system to enable real-time audio capture for downstream processing.

The Processing Module is implemented as a Python based desktop application that orchestrates three core functions: ASR, NMT, and TTS synthesis. The incoming audio stream is first transcribed to text using a speech recognition engine, after which the text is translated into a target language via Google's neural translation API. The translated text is then rendered into speech using a neural TTS engine.

The Output Module consists of a 5 x 5 parametric loudspeaker array that leverages ultrasonic transducers to project audio in a highly directional beam. An ESP32 microcontroller generates stable 40 kHz PWM signals, which modulate the ultrasonic carrier with the audio signal. This configuration enables focused sound projection, allowing the translated speech to a specific listener while minimising acoustic leakage into surrounding areas.

This tightly coupled modular architecture enables seamless interoperability between hardware and software components, supports real-time performance, and facilitates system scalability for future enhancements or deployments in varied operational contexts.

2.2. System Components

In this system, an 88 - 115 MHz FM wireless microphone is used for voice acquisition, paired with a standard FM receiver. This configuration facilitates untethered, low-latency voice input, which is well suited for mobile or spatially dynamic environments, with an adequate transmission range up to 100 meters. It ensures stable signal delivery with minimal power consumption.

The TPA3116D2 Class-D amplifier with its high-efficiency performance (over 90 %) supports both mono and stereo configurations, with an output power up to 50 W per channel. A total harmonic distortion plus noise (THD+N) rating of 0.1% ensures accurate reproduction of modulated audio signals, which is critical for maintaining intelligibility after demodulation in the air medium by the parametric speaker system.

The ESP32 microcontroller serves as the central control unit for PWM signal generation and timing synchronisation. Featuring dual-core processing, a 12-bit Analog-to-Digital Converter (ADC), and high-resolution PWM outputs, it is well suited for real-time signal modulation tasks. The ESP32 produces 40 kHz carrier frequency PWM signals required to drive the ultrasonic transducers, maintaining temporal precision and minimising phase jitter, to ensure high-fidelity directional audio synthesis.

The IR2184 MOSFET Driver is employed to control the high speed switching transients required for ultrasonic modulation. It provides isolated gate control with a floating high-side driver rated for high-voltage operation. These attributes are essential for efficient signal amplification in the ultrasonic driving circuit.

The IRLZ44N N-channel MOSFET is integrated into the output driving stage due to its low gate threshold voltage, high drain current capacity, and fast switching characteristics. These attributes make it well suited for high frequency operation in the ultrasonic transducer array, ensuring efficient and reliable modulation of the high voltage PWM signal without significant thermal losses or switching delays.

A parametric speaker is composed of 25 parallel ultrasonic piezoelectric transducers with the frequency range of 40 - 60 kHz is integrated to emit high frequency carrier waves that are modulated with the audio signal via PWM. Through nonlinear acoustic interactions in air (self-demodulation), these modulated signals reconstruct audible sound in a narrowly confined beam. Two such arrays are developed in this system to simultaneously broadcast translations in two distinct languages, each directed toward a different spatial zone. This configuration ensures spatial separation of audio output, listener specificity, and minimal acoustic interference in shared environments.

The desktop application implemented in Python, is structured into three principal functional modules: speech recognition, multilingual translation, and TTS synthesis. The speech recognition module utilises pre-trained STT models to transcribe live audio input into textual data with high accuracy. This transcription is then processed by the multilingual translation module, which leverages advanced NMT techniques to convert the source language text into the target language in text. Finally, the TTS module synthesises the translated text into natural, human like speech.

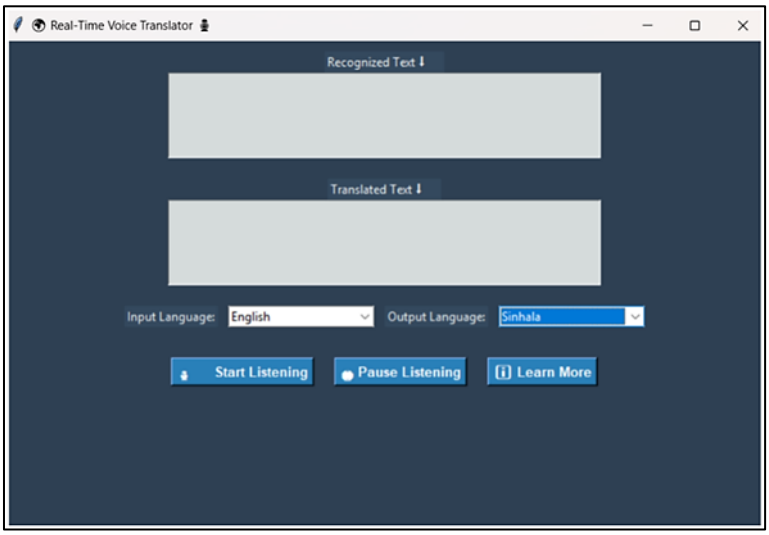

Figure 2 UI of the System

The synthesised audio is subsequently modulated into a PWM signal compatible with ultrasonic carrier frequencies for parametric speaker output. The entire processing pipeline is optimised for low latency, maintaining a balance between translation accuracy and real-time responsiveness. The user interface (UI) facilitating these operations is depicted in Figure 2.

Effective signal conditioning is essential to preserve audio fidelity during ultrasonic modulation and subsequent acoustic projection. The LM358 operational amplifier is employed to amplify and precisely shape the audio signals prior to the driver stage. The passive components, including resistors, capacitors, and protection diodes such as the UF4007 and Schottky 1N5819, are integrated within the signal path. These components perform critical functions including filtering unwanted noise, voltage regulation, frequency tuning, and protecting the circuitry against voltage spikes. This conditioning ensures the modulation signals delivered to the MOSFET driver and ultrasonic transducers are clean,

stable, and accurately represent the intended audio waveform, thereby maximising the clarity and effectiveness of the directional audio output.

3. Results and Discussion

The developed AI-based wireless real-time voice translation system successfully fulfills its core objectives by integrating instantaneous multilingual translation with spatially directed audio output. This solution demonstrates substantial potential for facilitating seamless communication in multilingual contexts. Although the experimental setup limited audio output channels to two target languages; Sinhala and Tamil, from an English source due to hardware constraints, the system architecture supports any language compatible with Google's translation API. The overall system design is illustrated in Figure 3.

Figure 3 Overall design of the System

The wireless FM microphone subsystem provided reliable and stable voice acquisition, enabling uninterrupted speech capture and transmission to the processing unit. Empirical evaluation of the Python based speech recognition module revealed an accuracy range between 90% and 95%, contingent upon ambient noise conditions and speaker enunciation clarity. The integrated NMT engine exhibited robust performance, maintaining a real-time processing latency between 1.5 and 2.5 seconds. This latency threshold is within acceptable bounds for interactive, dynamic conversational settings. The graphical user interface (Figure 2.) enhanced user experience by allowing intuitive language selection and monitoring of live transcription and translation.

The resultant acoustic beamforming achieved a highly directional audio projection, confining translated speech within a narrowly defined spatial zone. This spatial selectivity significantly reduces auditory interference and noise pollution in shared environments, thereby preserving the privacy and clarity of communication for the intended listeners.

The system operating on a regulated 12 V, 3 A DC power supply, maintained stable electrical current delivery over prolonged operational duration. Thermal measurements and power profiling indicated no significant fluctuations or overheating.

4. Conclusion

This work successfully validates the feasibility of integrating real-time multilingual speech processing with spatially directed, wireless audio transmission. The synergy of a Python based AI driven translation pipeline, FM wireless voice input, and ultrasonic parametric audio output provides a robust platform for multilingual communication across diverse domains such as public venues, educational institutions, and professional conferences.

The incorporation of the ESP32 microcontroller for generating 40 kHz - PWM signals facilitates precise modulation of ultrasonic transducers, thereby ensuring that translated speech is delivered with high clarity and minimal interference.

Key technical challenges including ambient noise affecting speech recognition, translation latency, and ultrasonic beam dispersion were addressed through comprehensive signal conditioning, software optimisation, and refined hardware

design. The resulting system demonstrates consistent translation accuracy of 90–95%, low end-to-end latency of 1.5 - 2.5 seconds, and operational stability under continuous use.

Looking ahead, there is considerable potential for system enhancement. Although the current prototype supports a single audio output channel due to hardware constraints, the software architecture is capable of parallel multi-language translation. Future work could extend this capability by employing stereo audio output to simultaneously deliver two distinct languages on separate channels. Furthermore, integrating external audio interfaces such as multi-channel sound cards or mixers could enable concurrent broadcasting of multiple languages, thereby broadening applicability to larger, more linguistically diverse audiences.

Compliance with ethical standards

Acknowledgments

We would like to express our sincere gratitude to the Department of Physical Science, Faculty of Applied Science, Trincomalee Campus, Eastern University, Sri Lanka for their invaluable support, including partial financial assistance and provision of laboratory facilities for this research project.

Disclosure of conflict of interest

The authors declare that there are no conflicts of interest or competing interests related to the publication of this manuscript.

References

[1] Bergen, A. D., et al. (1998). Wireless Audio Technologies: A Review. Journal of Audio Engineering.

[2] Chin, G. (2013). Real-Time Translation Systems in Education. Educational Technology Journal.

[3] Deng, L., & Li, X. (2013). Recent Advances in Speech Recognition and Language Processing. IEEE Transactions on Speech and Audio Processing.

[4] Espressif. (2020). ESP32 Technical Reference Manual. Espressif Systems.

[5] Iida, M., et al. (2009). Advances in Parametric Speaker Technology. Ultrasonics Symposium.

[6] Jones, M. (2005). Parametric Sound and Its Applications. Audio Engineering Society.

[7] Kraus, A., & Geng, X. (2016). Advancements in Wireless Audio Transmission Technologies. IEEE Transactions on Consumer Electronics.

[8] Koehn, P. (2009). Statistical Machine Translation. Cambridge University Press.

[9] Kuo, S., & Lee, W. (2010). Digital Signal Processing and Applications. Wiley-IEEE Press.

[10] Lin, H., et al. (2019). Efficient PWM Signal Generation for Audio Systems. Journal of Signal Processing.

[11] Maclean, A., et al. (2015). Multilingual Conference Systems: A Review of Technologies. Conference Proceedings.

[12] Mikulovic, J., et al. (2017). Translation Accuracy and Latency in Real-Time Systems. Proceedings of the International Conference on Machine Translation.

[13] Sato, M., et al. (1995). Development of Parametric Speakers Using Ultrasonic Waves. Journal of the Acoustical Society of Japan.

[14] Vasilenko, M., et al. (2017). Assistive Technologies in Multilingual Environments. Assistive Technology Journal.

[15] Vaswani, A., et al. (2017). Attention is All You Need. Advances in Neural Information Processing Systems.

[16] Xie, L., et al. (2019). Advancements in Wireless Microphone Systems. IEEE Transactions on Audio, Speech, and Language Processing.