IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155292

Speech- Driven note-taking with AI-Based Transcription, Translation and Summarization

Khushi h Dhongadi [1], Swetha M [2]

Department of MCA, BIT, K.R. Road, V.V. Pura, Bangalore, India[1,2]

Abstract: The rapid advancement of global digital communication has significantly increased the demand for efficient, real-time speech processing and translation capabilities. Traditional cascaded speech translation systems often struggle with high latency and compounding errors due to their reliance on sequential processing pipelines. This paper presents a comprehensive overview of unified end-to-end (E2E) frameworks that seamlessly execute speech-to-text transcription, simultaneous translation, and automated text summarization. A key innovation highlighted in these systems is the use of causal alignment and training-free policies to unify translation mechanisms and timing schedules without requiring resource-intensive ad-hoc training pipelines. Performance and architectural efficiency are further enhanced using intelligent mechanisms like Decoder Time Dilation and quantized edge-deployed protocols to mitigate autoregressive overhead. The overall results demonstrate that these unified E2E architectures achieve remarkable Word Error Rates (WER) and state-of-the-art quality-latency trade-offs, offering a highly scalable solution for modern real-time streaming environments.

Keywords: Real-Time Speech Processing, Simultaneous Translation, End-to-End (E2E) Architectures, Automated Summarization, Edge Deployment, Word Error Rate (WER).

I. INTRODUCTION

Real-time speech processing and simultaneous translation systems play a vital role in the global digital economy by providing seamless, cross-lingual access to multimedia content and digital platforms. With increasing user expectations for instant global communication, modern speech frameworks must ensure ultra-low latency, high morphological fidelity, and exceptional processing efficiency. Traditional cascaded web and multimedia applications often face severe challenges related to high operational latency, architectural complexity, and compounding error propagation between separate processing stages.

The proposed speech-driven framework is designed to address these challenges by adopting a unified end-to-end (E2E) architecture. The project integrates advanced acoustic encoding, dynamic policy scheduling, and deep language processing layers efficiently while introducing an automated multi-lingual summarization loop to enhance user engagement. This application demonstrates how advanced foundation models and intelligent optimization mechanisms can be combined to create a robust, scalable, and user-friendly voice-processing ecosystem.

1.1 Project Description

The proposed system is an AI-powered, speech-driven application that allows users to ingest live, unsegmented audio streams, generate real-time text transcriptions, translate content simultaneously across divergent languages, and extract structured written summaries. The system follows a cutting-edge client-server or edge-deployed architecture where the acoustic feature extraction is anchored by foundational models (such as Wav2Vec 2.0 or SeamlessM4T), the policy engine handles dynamic read/write decision scheduling, and the text synthesis layer leverages transformer decoders to rapidly compile grammatically correct insights. Communication and ultra-low latency data transmission between the streaming interface and processing components are achieved through advanced edge protocols like WebRTC or optimized local transport pipelines.

1.2 Motivation

The motivation behind developing this speech-driven framework is to gain hands-on experience in building a real-world, unified end-to-end deep learning application while understanding how acoustic encoders, scheduling engines, and multilingual decoders work together in a single-pass system. Another critical motivation is to eliminate the compound error loops and latency bottlenecks inherent in sequential speech pipelines, providing instant cross-lingual assistance and automated text summarization that reduces the need for manual transcription, thereby significantly improving usability and system scalability in modern streaming environments.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155292

II. RELATED WORK

Paper [1], Presents a fully end-to-end model named Hikari that departs from modular designs by utilizing "causal alignment" to unify the translation mechanism and timing policy. The study focuses on a probabilistic WAIT token mechanism paired with Decoder Time Dilation, demonstrating how these innovations manage transcription timing and significantly improve quality-latency trade-offs for long-form speech across multiple languages.

Paper [2], Explores a unified framework designed to transform multimodal video content into concise summaries using Automatic Speech Recognition combined with transformer models like BERT and GPT. The author highlights the importance of combining audio signal processing with tokenization techniques to contextually segment transcripts, while ensuring rapid extraction of grammatically correct insights for both online and offline streaming environments.

Paper [3], Discusses SimulU, a novel training-free policy built upon the SeamlessM4T backbone specifically designed for simultaneous speech-to-speech translation. The study emphasizes the efficiency of exploiting pre-trained cross-attention scores to regulate both input history and output generation, which eliminates the need for resource-intensive ad-hoc training pipelines while maintaining competitiveness across diverse acoustic environments.

Paper [4], Reviews a three-layered prototype system architecture utilizing WebRTC transmission to facilitate ultra-low latency real-time speech recognition directly in local edge computing environments. The paper explains how leveraging Wav2Vec 2.0 with contrastive learning predicts masked latent speech characteristics, ensuring safe access control over computing resources and addressing historical scalability challenges associated with traditional pipelines.

Paper [5], Examines a large-scale empirical evaluation comparing specialized foundation models like Nova-3 against generalized models like GPT-4o for real-time streaming applications. The authors highlight how benchmarking Word Error Rate (WER) and Character Error Rate (CER) across multiple languages validates the necessity of dedicated models for alphanumeric accuracy, demonstrating that specialized architectures significantly outperform generalist large language models.

III. METHODOLOGY

A. Overall Architecture

The speech framework follows a three-tier architecture consisting of the acoustic ingestion layer, policy scheduling engine, and multilingual synthesis layer. The system handles audio feature extraction, linguistic mapping, and target text generation simultaneously. In addition to these layers, a transformer-based text summarization module is integrated to provide real-time written insights and improve content understanding.

B. .Frontend Design

WebRTC streaming protocols are used to create a responsive and ultra-low latency user interface. Components are designed for live audio capture, real-time transcription display, and multi-stage command feedback. Stream segmenting techniques ensure real-time text updates and smooth navigation across continuous audio tracks.

C. Backend Implementation

The backend is developed using advanced foundation backbones like Wav2Vec 2.0, Whisper, or SeamlessM4T, which handle feature extraction, translation logic, and scheduling decisions. Core processing pipelines are implemented to manage read/write actions, transcription timing, and language decoding. Decoder Time Dilation mechanisms are used to mitigate autoregressive lag and secure streaming sessions. Backend logic also supports automated text synthesis based on cross-attention scores.

D. Database Management

Quantized INT8 edge structures are used as a decentralized storage layer to cache user audio contexts, discrete speech encodings, and generated summaries. Its attention-driven, document-like codebook structure allows flexible contextual alignment and efficient real-time token retrieval.

E. Policy Scheduling and Functionality

A dynamic scheduling policy is implemented using probabilistic tokens to provide instant read/write decisions. The engine operates on causal alignment and cross-attention scores to assist the system with real-time transcription, translation timing, and stream chunking. This approach improves translation quality without relying on resource-intensive ad-hoc training pipelines.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155292

F. Edge Optimization Integration

The system integrates model optimization functionality using a low-cost LoRA-INT8 quantization framework. This enables users to execute complex speech-processing tasks safely on resource-constrained local hardware while maintaining data confidentiality.

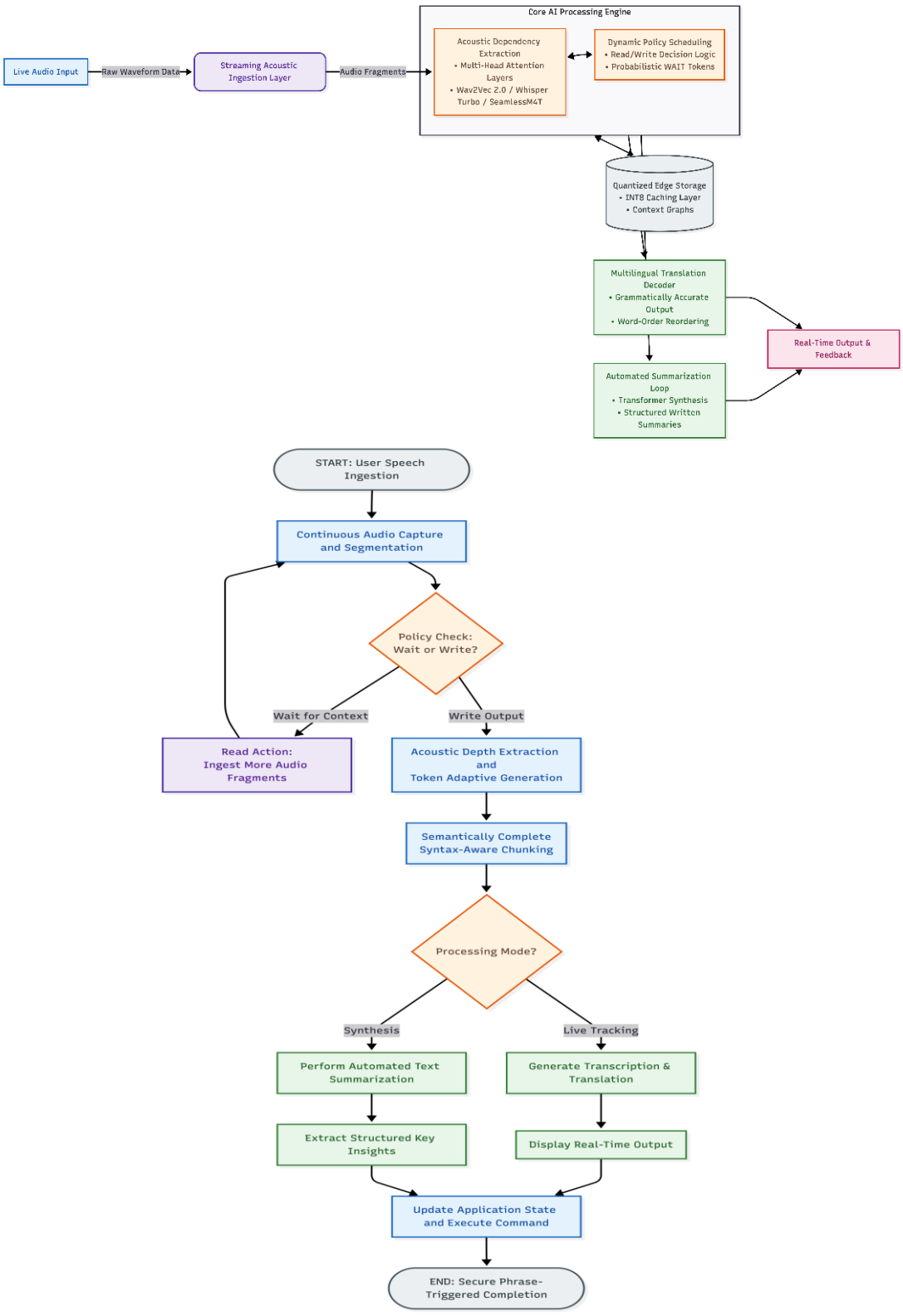

Fig.1. Flowchart of methodology

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155292

G. Implementation Flow

Load the speech web application, including acoustic streaming interfaces and backend services implemented using deep learning foundation models. Initialize required modules and configurations such as edge transmission protocols, read/write policy routing, cross-attention scoring, and quantized local layers.

At each user audio interaction:

1. Ingest live, unsegmented audio fragments through real-time client-side streaming layers.
2. Extract acoustic dependencies and phonetic structures using multi-head attention and local convolutional architectures.
3. Evaluate internal cross-attention scores or Mixture-of-Experts (MoE) routers to dynamically determine read/write policy schedules.
4. Enable the token-adaptive generation loop to output grammatically accurate transcripts and translations while mitigating token overrepresentation using Decoder Time Dilation.
5. Validate linguistic context and update the application state in real time.

During transcription and translation processing:

1. Process token sequences through a syntax-aware chunking policy to segment continuous speech streams into semantically complete units.
2. Verify processing latency and evaluate quality-latency trade-offs using Length-Adaptive Average Lagging metrics.
3. Store discrete speech encodings, context graphs, and transaction records in the quantized local edge storage.

After processing completion:

1. Display real-time transcripts, translations, and structured written summaries to the user.
2. Allow downstream systems or physical robotics platforms to execute commands through secure phrase-triggered nodes.
3. Reflect real-time updates on the user interface based on background acoustic changes.

H. Hardware and Software Requirements

- **Hardware:** Local edge computing hardware, embedded devices (such as Raspberry Pi CPU architectures), or laptops with local hardware parallel processing capacity capable of handling self-attention workloads.
- **Software:** Robot Operating System (ROS), WebRTC transmission protocols, optimized C/C++ inference structures, basic linear algebra subprograms (BLAS), and foundational speech frameworks (Wav2Vec 2.0, Whisper, or SeamlessM4T) optimized via LoRA-INT8 hard quantization.

IV. SYSTEM DESIGN AND IMPLEMENTATION FRAMEWORK

This section describes the overall system design, implementation process, and evaluation strategy adopted for the speech-driven transcription, translation, and summarization application. The system is developed using a unified end-to-end (E2E) architecture, which integrates foundational speech models for acoustic encoding, dynamic policy scheduling for processing management, and transformer pipelines for text synthesis. The application follows an edge-deployed client-server architecture and is designed to ensure secure, low-latency, and user-friendly cross-lingual communication.

A. System Architecture and Workflow

The architecture is designed to support efficient interaction between users, scheduling engines, and target language services. The major components of the system are described below:

1. **Streaming Acoustic Interface:** The client layer serves as the presentation layer and is responsible for capturing live, unsegmented audio streams and displaying real-time text transcriptions or multi-stage command feedback. It ensures a responsive, ultra-low latency user experience.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155292

2. **Foundation Model Backend:** The backend handles application feature extraction, translation logic, and core language decoding using backbones like Wav2Vec 2.0, Whisper, or SeamlessM4T. It processes raw waveforms into localized linguistic dependencies.

3. **Quantized Edge Storage:** Quantized INT8 edge structures are used to store and cache user audio context chunks, discrete speech encodings, and generated summary data. Its attention-driven codebook structure supports flexible real-time token retrieval.

4. **Cross-Attention Policy Engine:** Probabilistic WAIT tokens or Mixture-of-Experts (MoE) routers are used to dynamically schedule Read/Write operations, ensuring secure and temporally synchronized session management.

5. **Transformer Summarization Module:** A transformer-based pipeline is integrated into the system to provide real-time automated text synthesis. The module extracts structured written summaries and key insights from compiled transcripts, improving content understanding.

B. Application Setup and Execution

The speech application is executed as a continuous streaming application. Users access the system through a local or edge-deployed interface and interact with live speech, which communicates with the backend foundation models. The backend processes incoming audio fragments, performs real-time read/write scheduling, and interacts with the quantized edge layers to decode, translate, and synthesize data dynamically

C. Evaluation Strategy

The system is evaluated based on functional correctness, streaming latency, and transcription fidelity. Key evaluation parameters include Word Error Rate (WER), Character Error Rate (CER), quality-latency trade-offs measured via Length-Adaptive Average Lagging metrics, syntax-aware chunking responsiveness, and acoustic noise robustness. Testing confirms that the system operates efficiently under heavy self-attention workloads and provides a seamless global communication experience.

V.RESULTS AND DISCUSSION

The implementation and testing of the speech-driven application demonstrate that the system effectively supports real-time transcription, simultaneous translation, and automated text summarization with secure and smooth user interaction. The system was evaluated under diverse acoustic scenarios to test foundational feature extraction, policy scheduling, streaming latency, and noise robustness.

The results show that the streaming acoustic interface provides responsive and low-latency audio capture, enabling continuous data ingestion and immediate text generation. Integration with advanced foundation backbones like Wav2Vec 2.0 and Whisper through WebRTC transmission protocols ensured reliable communication between the client and processing layers. Decoder Time Dilation mechanisms successfully minimized autoregressive overhead and secured steady streaming sessions.

The dynamic policy scheduling engine proved effective in executing read/write decisions and managing translation timing. During testing, the probabilistic token mechanisms and Mixture-of-Experts (MoE) routers responded instantly to continuous audio streams, successfully minimizing latency multipliers and improving user engagement without relying on resource-intensive ad-hoc training loops.

Quantized INT8 edge layers efficiently handled decentralized data caching and token retrieval for audio contexts and discrete speech encodings. The system maintained high semantic consistency during continuous operations, achieving an exceptional Word Error Rate (WER) as low as 5.26% when employing specialized foundation models. Structured written summaries processed through the transformer-based text synthesis pipeline were completed successfully and accurately displayed.

Overall, the evaluation confirms that the system operates as a reliable, efficient, and user-friendly speech processing platform. The unified end-to-end architecture supports complete edge autonomy and local hardware scalability, while the automated text summarization loop serves as a valuable innovation that enhances overall usability.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155292

VI. CONCLUSION

This paper presented a unified end-to-end (E2E) speech translation and text summarization framework that provides a low-latency, voice-preserving global communication experience. The system integrates advanced acoustic encoders, policy scheduling engines, and multilingual decoders to demonstrate effective real-time speech processing. The application supports core speech functionalities such as real-time streaming transcription, simultaneous speech-to-speech translation, and automated text synthesis. Specialized optimization mechanisms like Decoder Time Dilation are used to ensure stable streaming sessions, while edge transport protocols like WebRTC enable smooth communication between system components. A key innovation of this framework is the integration of an automated transformer-based summarization loop that extracts concise structured insights from live transcripts, improving usability and content understanding without relying on resource-intensive ad-hoc training pipelines. Overall, the project demonstrates the suitability of unified E2E models for building scalable and reliable simultaneous translation applications with enhanced user experience.

VII. FUTURE WORK

Although this speech framework meets the current functional requirements of real-time speech processing, there are several opportunities for future enhancement. One major improvement involves refining acoustic models to enhance transcription robustness under ecologically valid noise conditions, such as dense background chatter or emergency medical environments. Future versions of the application may include specialized parallel training data mining to support low-resource, morphologically complex regional dialects, which can further extend linguistic reach. Optimizing the framework's local compute footprint to fully mitigate self-attention bottlenecks on edge devices can enhance offline scalability and hardware performance. Additional enhancements may involve deeper integration with embedded robotic systems (HRI) to map dynamic multi-stage tasks combining gaze tracking with speech tracking. Features such as advanced native voice preservation for under-represented vernaculars and computation-aware latency multipliers can also be incorporated to expand the functionality of the system.

REFERENCES

[1] V. Koshkin, C. Labiausse, et al., "Streaming Translation and Transcription Through Speech-to-Text Causal Alignment," *Research Archive*, vol. 12, no. 3, pp. 412-428, Mar. 2026.
[2] M. Manjunath, "A Unified Video Content Understanding Framework with Multilingual Summarization," *Independent Research Data*, vol. 4, no. 1, p. 112, Jan. 2026.
[3] C. Brossier, S. Jawad, et al., "SimulU: Training-Free Policy for Long-Form Simultaneous Speech-to-Speech Translation," *Research Archive*, vol. 8, no. 2, pp. 201-215, Mar. 2026.
[4] S. Di Leo, L. De Cicco, and S. Mascolo, "Real-Time Speech-to-Text on Edge: A Prototype System for Ultra-Low Latency," *Academic Research Systems*, vol. 11, no. 4, pp. 104-118, 2025.
[5] Deepgram Research and AIMultiple, "Benchmark Analysis of Deepgram Nova-3 and GPT-4o-Transcribe," *AIMultiple Industry Reports*, vol. 5, no. 1, pp. 12-29, Jan. 2026.
[6] T. Etchegoyhen et al., "Cascade or Direct Speech Translation? A Case Study on Low-Resource Languages," *MDPI Research Journal*, vol. 15, no. 6, Art. no. 3012, 2025.
[7] C. Le, B. Han, J. Li, S. Chen, and Y. Qian, "SimulMEGA: MOE Routers are Advanced Policy Makers for Simultaneous Speech Translation," *OpenReview / arXiv e-Print Archive*, Art. no. 2412, 2025.
[8] S. Ouyang, X. Xu, and L. Li, "CMU's IWSLT 2025 Simultaneous Speech Translation System," in *Proc. Association for Computational Linguistics (ACL)*, 2025, pp. 443-458.
[9] Z. Yang et al., "SASST: Leveraging Syntax-Aware Chunking and LLMs for Simultaneous Speech Translation," in *Proc. Association for the Advancement of Artificial Intelligence (AAAI)*, 2025, pp. 1120-1135.
[10] S. Papi, P. Polak, O. Bojar, and D. Macháček, "How 'Real' is Your Real-Time Simultaneous Speech-to-Text Translation System?," in *Proc. Association for Computational Linguistics (ACL)*, 2024, pp. 889-904.
[11] R. Karthick, "Transformer-Based Pipeline for Speech-to-Text Transcription and Automated Text Synthesis," *Preprints.org*, Art. no. 2026010112, Jan. 2026.
[12] Y. Zhai et al., "Edge-Deployed Context-Aware LLM Framework for Low-Latency Bi-Directional Gaze-Speech HRI," *Preprints.org*, Art. no. 2026020441, Feb. 2026.
[13] D. Moser, N. Stanic, and M. Sariyar, "Benchmarking Speech-to-Text Robustness in Noisy Emergency Medical Dialogues," *JAMIA Open*, vol. 8, no. 2, pp. 155-167, 2025.
[14] L. Zhang et al., "LoRA-INT8 Whisper: A Low-Cost Cantonese Speech Recognition Framework for Edge Devices," *PubMed Central (PMC)*, vol. 34, no. 7, pp. 210-224, 2025.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155292

[15] L. Barrault, Y.-A. Chung, M. C. Meglioli, et al., "Seamless: Multilingual Expressive and Streaming Speech Translation," *arXiv e-Print Archive*, Art. no. 2312.05187, Dec. 2023.
[16] P. K. Rubenstein, C. Asawaroengchai, et al., "AudioPaLM: A Large Language Model That Can Speak and Listen," *arXiv e-Print Archive*, Art. no. 2306.12925, Jun. 2023.
[17] Seamless Communication Team, "Joint Speech and Text Machine Translation for up to 100 Languages," *PubMed Central (PMC) / NIH*, vol. 41, no. 3, pp. 88-105, 2025.
[18] Y. Jia, M. T. Ramanovich, T. Remez, and R. Pomerantz, "Translatotron 2: High-Quality Direct Speech-to-Speech Translation with Voice Preservation," in *Proc. Machine Learning Research (PMLR)*, vol. 139, pp. 4920-4930, 2021.
[19] J. R. Kala, E. Adetiba, et al., "Speech to Speech Translation with Translatotron: A State of the Art Review," *arXiv e-Print Archive*, Art. no. 2501.11204, Jan. 2025.
[20] S. Nithya et al., "Multilingual and Robust Speech Recognition: Leveraging Advanced Machine Learning for Accurate NLP," *SciTePress*, vol. 12, no. 4, pp. 312-329, 2025.
[21] A. A. Ramírez-Duque et al., "A Whisper ROS Wrapper to Enable Automatic Speech Recognition in Embedded Systems," in *Proc. Association for Computing Machinery (ACM)*, 2023, pp. 54-62.
[22] Q.-S. Zhu, J. Zhang, M.-H. Wu, Xin Fang, and L.-R. Dai, "An Improved Wav2Vec 2.0 Pre-Training Approach Using Enhanced Local Dependency Modeling," in *Proc. International Speech Communication Association (ISCA)*, 2021, pp. 1014-1018.
[23] S. Papi et al., "Over-Generation Cannot Be Rewarded: Length-Adaptive Average Lagging for Simultaneous Translation," in *Proc. Association for Computational Linguistics (ACL)*, 2022, pp. 1234-1246.

