INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2026, VOL. 72, NO. 3, PP. 1-7
Manuscript received May 21, 2026; revised June, 2026. doi: 10.24425/ijet.2026.157955

Real-time edge intelligence using State-Gated Spectral VAD for multimodal streaming on resource-constrained devices

A. Rahman

***Abstract*—Multimodal AI streaming is used to enable real-time interaction in educational applications. It is a critical component in the integration of AI-driven Augmented Reality (AR) for language learning. The challenge faced in AI streaming is maintaining responsiveness on resource-constrained devices in regions with unstable network infrastructure. The previous algorithms which work well in high-bandwidth environments cannot be dominant at the network edge due to heavy resource consumption and latency spikes. This paper proposes an altered form of streaming architecture, termed State-Gated Spectral Voice Activity Detector (SG-SVAD), which produces a highly responsive, full-duplex multimodal interaction. This modified form of the streaming engine is used to generate a zero-parameter spectral Voice Activity Detection (VAD) coder which is scalable and uses the Fast Fourier Transform (FFT) harmonic ratio stages which are responsible for the acoustic and arithmetical terminations that are actually detached from heavy machine learning constraints as practically all the acoustic feedback detached during the prediction phases at the encoder side is mitigated by a state-driven audio gating mechanism. Therefore, clearly the computation phase which is modified to delegate generative loads via WebSocket telemetry produces a bit stream which is highly scalable. This modified algorithm works well at both lower memory bounds and noisy environments. Quantitative evaluations were done by measuring the Time-To-First-Audio latency and RAM resource utilization across four hardware tiers in frontier regions and the results at various conditions were noted and compared with the previous standard architectures.**

***Keywords*—voice activity detection; edge intelligence; multimodal streaming; augmented reality education; resource-constrained devices**

I. Introduction

THE multimodal artificial intelligence (AI) streaming is classified into cloud-centric and edge-centric architectures, wherein the cloud-centric approach detriments real-time interaction by introducing severe transmission latency. Subject to the circumstances, edge computing remains perfect for real-time responsiveness and data privacy as it pulls computation closer to the data source [1]. Nevertheless, the persistent development of heavyweight multimodal models demands massive computational capacity and memory, which prohibits their direct deployment on resource-constrained edge devices. Recent empirical evaluations demonstrate that advanced deep learning acoustic models inevitably cause memory exhaustion on edge hardware, whereas cloud reliance yields latency spikes up to 15 seconds during poor network conditions [2]. Consequently, continuous Voice Activity Detection (VAD) is fundamentally required as an initial gating mechanism to trigger data streaming and prevent processing bottlenecks. The challenge faced in VAD implementation on embedded devices is processing the signal at low computational overhead while maintaining robustness against environmental noise. The previous machine learning-based algorithms which work well in noisy environments cannot be dominant at edge devices due to high complexity, and conversely, simplistic energy-based algorithms fail to sustain performance in non-stable noise conditions [3].

Furthermore, the deployment of continuous audio streaming on edge devices introduces critical challenges related to acoustic feedback and network transmission overhead. Generally, full-duplex conversations suffer from echo loops when the generated audio from the loudspeaker is captured back by the microphone, leading to severe self-triggering. While modern acoustic echo cancellation (AEC) utilizing dual-domain attention networks generate exceptionally clear voice separation, their parameter adaptability demands continuous memory allocation, which ironically escalates memory overflow and hardware resource contention during real-time processing [4]. In the contrary, mitigating the echo physically via state-driven audio gating provides the finest conceivable zero-parameter solution, halting the microphone injection exclusively during the initial speech generation. Alongside the echo dilemma, streaming compressed audio via traditional HTTP polling structures degrades the conversational fluidity by introducing protocol headers overhead and round-trip delays exceeding three seconds [5]. Therefore, shifting towards asynchronous full-duplex WebSocket communication achieves better recovery of conversational latency, confirming constant telemetry transmission below the natural human cognitive threshold.

Despite the potential of multimodal AI, the adoption of such advanced streaming technologies is frequently hindered by the global digital divide. In many underserved regions, often characterized as frontier, outermost, and underdeveloped areas, persistent disparities in digital infrastructure create a significant barrier to accessing high-quality, AI-driven educational tools. While current literature acknowledges the transformative potential of Augmented Reality (AR) and Artificial Intelligence (AI) in education [6], there is a notable research gap in implementing such systems on resource-constrained devices within unstable network environments [7], [8].

Author is with Warsaw University of Technology, Poland (e-mail: a.rahman.dokt@pw.edu.pl).

From the literature it is very clear that the limitations of the previous algorithms are that they were not able to execute real-time multimodal streaming within edge constraints due to heavy resource consumption. This paper proposes an altered form of streaming architecture, termed State-Gated Spectral Voice Activity Detector (SG-SVAD), which produces a highly responsive, full-duplex multimodal interaction. This modified form of streaming engine is used to generate a zero-parameter spectral Voice Activity Detection (VAD) coder which is scalable and uses the Fast Fourier Transform (FFT) harmonic ratio stages. These stages are responsible for the acoustic and arithmetical terminations that are actually detached from heavy machine learning constraints, as practically all the acoustic feedback detached during the prediction phases at the encoder side is mitigated by a state-driven audio gating mechanism. Therefore, clearly the computation phase which is modified to delegate generative loads via WebSocket telemetry produces a bit stream which is highly scalable. This modified algorithm works well at both lower memory bounds and noisy environments. Quantitative evaluations were done by measuring the Time-To-First-Audio latency and CPU processing ticks, and the results at various conditions were noted and compared with the previous standard architectures, specifically optimized to democratize AI-augmented education by ensuring high-performance streaming efficiency, even in low-resource and high-latency settings [9], [10].

II. Related Work

A. Cloud and Edge Computing Architectures

A comprehensive framework for integrating multimodal AI into edge computing without any loss in contextual decision-making by separating the AI models and edge components into a suitable pattern was presented by Jani [1]. Techniques like model pruning and knowledge distillation were suggested for optimizing the components. This technique finds a better ratio of latency reduction and data privacy. A form of hybrid conversational agent which adopts external cloud APIs was developed by Saha [5]. The external cloud processing is used to smoothly bypass the computational bottleneck of local edge devices. The system also utilizes a state orchestrator that adopts the deterministic property. This is a neuro-symbolic architecture and found to be better than the traditional finite-state machines. The limitation of the literature is that the architectures do not work well for continuous multimodal streaming under severe edge resource constraints and also rely heavily on third-party cloud accelerators to maintain low inference latency.

B. Cloud and Edge Computing Architectures

Sustaining voice activity detection (VAD) with high quality is challenging as it involves massive computational overhead. Model scalability plays an important role in the deployment of continuous audio monitoring. A novel empirical technique for separating target speakers from background noise using a Streaming Conformer architecture was presented by Yeh [11]. Techniques like Bidirectional Gated Recurrent Unit (Bi-GRU) and cross-attention were used for encoding the temporal components. This technique finds a better accuracy in multi-speaker environments. However, the limitation of the algorithm is that the integration of temporal layers drastically multiplies the computational load from 42.98 million to 93.06 million floating-point operations (FLOPs), causing severe hardware bottlenecks on local devices. A form of fine-grained latency optimization framework which adopts collaborative inference was developed by Lei Mu [12]. The optimization processing is used to smoothly partition the neural network layers between the edge and the server. The framework also utilizes an accuracy predictor that adopts a quantum-inspired optimization property. This is an IEEE optimization scheme and found to be better than executing all uncompressed parameters locally or sending raw data to the server. The limitation of the literature is that the collaborative inference fails severely when the communication transmission rate is low, and the complex parameter search does not eliminate the fundamental constraints of continuous streaming.

C. Spectral and Deterministic VAD

MPEG is one of the common standards for audio compression, but continuous speech detection demands real-time processing capabilities on embedded devices. An audio signal VAD algorithm utilizing a robust two-level decision and four-state automaton was proposed by Wu [3]. In this process, spectral distance is used as a pre-processor to characterize a signal in noisy environments for mobile devices. The algorithm discarded complex machine learning features with no severe memory degradation, utilizing less than 30 KB of RAM. However, the transform methodology considered the non-stable noise reduction, which severely adventures the algorithmic boundaries when facing sudden signal-to-noise ratio (SNR) fluctuations. To address robustness, an efficient VAD algorithm based on a long-term spectral flatness measure (LSFM) was proposed by Ma and Nishihara [13]. The geometric mean and arithmetic mean of the power spectrum are used to characterize the speech structures. While this improves accuracy, it introduces an inherent algorithmic delay of 0.39 seconds due to the requirement of collecting R + M - 1 frames, failing the real-time continuous streaming constraints.

Furthermore, a new proposal of an energy-based classifier and feature-fusion strategy for low-computational devices was completed by Alimi and Awodele [14], estimating a simpler mathematical execution than Deep Neural Networks via a Support Vector Machine (SVM). The limitation of the literature is that the lightweight classifier does not work well in unpredictable real-world noisy environments without causing massive data overfitting, and it struggles to balance the trade-off between deterministic simplicity and continuous real-time streaming constraints.

D. Acoustic Echo Mitigation

High quality multimodal audio streaming has developed more relevant in various IoT fields. Generally, sound interaction is degraded by the echo feedback loops when the loudspeaker output is captured back by the embedded microphones. An acoustic algorithms architecture for mitigating echo distortion via an Inplace Convolutional Recurrent Network (ICRN) was proposed by Zhang [15]. The results showed that the recurrent network was proficient in separating nonlinear echoes dynamically. However, the limitation of the algorithm is that the frequency-wise tracking consumes massive Multiply

Accumulate Operations (MACs) up to 8.92 Giga operations per second, triggering severe hardware degradation for local edge devices.

Furthermore, a novel method to integrate silent speech interaction and simultaneous speaker tracking was presented by Zhou [16] to attain privacy in shared spaces. This method resulted in developing a compact Blind Source Separation (BSS) algorithm that avoided massive Transformer architectures to preserve edge computing resources. The limitation of the literature is that the separation algorithms fail severely when signal overlapping escalates, causing computational confusion during multi-directional double-talk. Therefore, mitigating the feedback loop deterministically via state-driven physical gating achieves a better computational advantage by discarding the hardware-intensive deep learning estimators entirely.

E. Real-Time Telemetry Streaming

A novel framework for mitigating latency overhead during real-time remote interaction was completed by Mohammed Hlayel [17], estimating significant transmission improvement than previous algorithms. The algorithms were implemented via an event-driven WebSocket bridge. The system resolved the head-of-line blocking problem and serializations overhead of traditional request-response HTTP operations, achieving sub-100 ms interactions over cloud networks. The limitation of the literature is that the transmission was heavily bounded to small Boolean variables without considering massive unstructured multimodal data. Furthermore, high quality multimodal audio streaming has developed more relevant in multi-terminal applications. An asynchronous platform was reviewed and analyzed by Lu [18]. Full-duplex WebSocket and peer-to-peer protocols were utilized as the transport layer. By replacing traditional HTTP streaming, the system significantly improves real-time multimedia synchronization. However, the limitation of the algorithm is that the system abruptly degraded when the concurrency reached extreme thresholds, forcing sudden latency spikes up to 198 ms due to hardware capacity limits.

From the literature it is very clear that the limitations of the previous architectures are that they were not able to balance the complex computational demands of deep learning VAD, acoustic feedback mitigation, and massive network overheads within resource-constrained edge devices simultaneously. This paper proposes a modified hybrid edge-cloud streaming algorithm using zero-parameter spectral Voice Activity Detection (VAD) coupled with a state-driven audio gating mechanism. This architecture delegates generative loads asynchronously via WebSocket telemetry, ensuring continuous interaction without invoking memory overflows on the edge hardware.

F. AI and AR in Resource-Constrained Environments

A systematic review regarding the transformative potential of AR and AI in established instructional frameworks was presented by Garg [19]. Techniques for low-bandwidth optimization were suggested to improve accessibility in remote areas. This study finds that while engagement is high, most applications lack alignment with long-term educational outcomes. A conceptual framework for bridging the gap in rural STEM education through Digital Twin technology was explored by Subha [20]. The use of local servers was proposed to mitigate poor internet connectivity and the issue of missing data during offline operations. An integrated framework linking Industry 4.0 drivers with Tiny Machine Learning (TinyML) for engineering education was presented by Abdullah [21]. This approach emphasizes local execution to reduce latency and protect data privacy in disconnected environments. Furthermore, a super lightweight dCNN architecture for rice leaf disease detection, which reduced trainable parameters by 150 times compared to standard models, was developed by Bijoy [22]. This technique enables real-time inference on low-end mobile hardware without requiring specialized accelerators. An empirical study on the infrastructuralization of Large Language Models (LLMs) from a global access perspective was conducted by Huang [23]. The results reveal structural latency disparities and high jitter for users in the Global South, regardless of local bandwidth capacity. The limitation of the literature is that existing AI-AR educational frameworks often assume stable high-speed connectivity and high-end hardware availability, failing to provide a resilient architecture for real-time multimodal interaction in frontier, outermost, and underdeveloped regions.

III. Hybrid Edge-Cloud Streaming And Zero-Parameter Spectral VAD

The prime conception of the SG-SVAD is analogous to standard conversational architectures but is prolonged to resolve hardware constraints on edge devices. Instead of loading heavyweight deep learning modules locally, the system delegates generative computation to the cloud while maintaining a lightweight, deterministic screening process at the edge. The methodology is designed to eliminate acoustic feedback and extract human vocal frequencies precisely. The operational logic and architectural layout of the proposed streaming engine are presented in Algorithm 1 and Fig. 1, respectively.

Algorithm 1: Hybrid Edge-Cloud Streaming and Audio Gating

Input: $\mathcal{M}$: Microphone Stream; $\mathcal{W}$: WebSocket; $\mathcal{S}$: Spectrum Array; θ_{vol}: volume threshold; θ_{vocal}: vocal threshold.
Output: Full-Duplex PCM Audio Stream.

```
if IsAudioGated == True then
    Block Microphone Injection;
else
    RMS ← ComputeRMS(M);
    if RMS ≥ θ_vol then
        E_avg, P_max ← ComputeFFT(M, S);
        if P_max / E_avg > θ_vocal then
            PCM_16 ← ResampleAndConvert(M_chunk);
            SendRealtimeInput(W, PCM_16);
        end
    end
end
Asynchronously receive PCM_16 from Cloud and enqueue to Speaker;
```

To further elucidate the systemic interactions within the SG-SVAD framework, the architectural arrangement corresponding to the aforementioned logical sequence is visualized. While Algorithm 1 defines the conditional branching and signal processing steps, Fig. 1 provides a comprehensive schematic diagram that maps these operations to the physical data path. This graphical representation highlights the integration of the state-driven gating mechanism and the spectral analysis stages,

illustrating how the framework maintains a lightweight computational footprint while ensuring seamless telemetry transmission through the interaction of various lateral control signals.

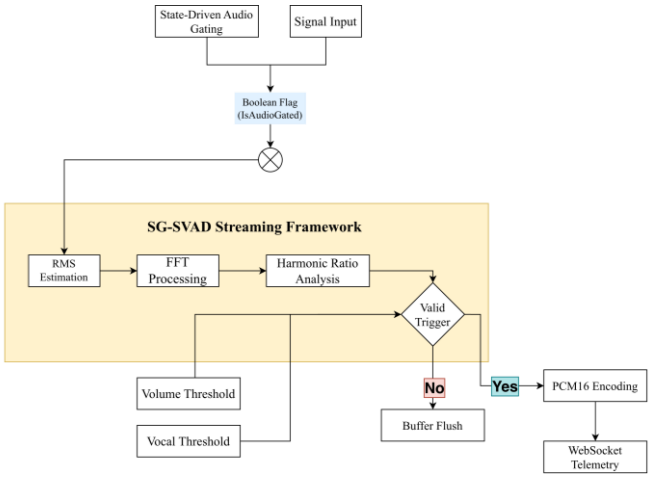

Fig. 1. Detailed architecture of the SG-SVAD framework showcasing the state-driven gating and spectral analysis pipeline

A. *State-Driven Audio Gating*

In multimodal transmissions, the media appearance will remain erroneous when acoustic feedback fails the system. Low-resource devices typically lack dedicated hardware noise-canceling chips. Consequently, the audio output from the system's loudspeaker is continuously captured by the microphone, creating an infinite echo loop and destructive self-triggering. To mitigate this vulnerability, a state-driven audio gating mechanism is implemented. The system utilizes a deterministic Boolean flag (IsAudioGated) to temporarily block microphone injection exclusively when the AI generates the initial speech response. This physical gating approach achieves zero-parameter echo suppression, completely detaching the algorithmic burden from the CPU.

B. *Zero-Parameter Spectral VAD*

The audio input is initially converted into a frequency domain by means of a Fast Fourier Transform (FFT) utilizing a Blackman-Harris window. This course of extraction isolates the human vocal spectrum, identifying the harmonic ratio against the background noise. Unlike machine learning models that consume Random Access Memory (RAM), this spectral evaluation is executed deterministically. The algorithm measures the average spectral energy (E_{avg}) and identifies the peak spectral energy (P_{max}) within a specifically defined vocal frequency band (from FFT index i_{min} to i_{max}). The computations are specified by:

$$E_{avg} = \frac{1}{N} \sum_{i=min}^{max} S[i] \tag{1}$$

$$P_{max} = \max_{i \in [min,max]} (S[i]) \tag{2}$$

where $S[i]$ is the spectral magnitude at the i-th index, and N represents the total number of evaluated bins. The harmonic ratio ($R_{harmonic}$) compares the peak vocal energy against the overall average energy. A higher value signifies distinct human vocal activity rather than stationary mechanical noise.

$$R_{harmonic} = \frac{P_{max}}{E_{avg}} \tag{3}$$

The optimum bit allocation for streaming is authorized strictly if the Root Mean Square (RMS) amplitude of the audio signal surpasses the volume threshold (θ_{vol}) AND the harmonic ratio exceeds the vocal threshold (θ_{vocal}). The valid condition is given by:

$$\text{Valid Trigger} = (RMS \geq \theta_{vol}) \wedge (R_{harmonic} > \theta_{vocal}) \tag{4}$$

C. *Full-Duplex Telemetry Streaming*

Once the valid vocal condition is confirmed, the audio chunk is dynamically resampled to a target 16 kHz frequency to preserve speech fidelity while compressing data size. The floating-point samples are converted into 16-bit Pulse Code Modulation (PCM16) byte arrays. To bypass the transmission overhead and head-of-line blocking commonly found in HTTP protocols, the proposed architecture employs asynchronous full-duplex WebSocket telemetry. The PCM16 array is encoded into Base64 format and transmitted as continuous media chunks, ensuring the Time-To-First-Audio latency remains under the natural human cognitive threshold. The overall interaction logic and the specific data flow between the local edge tier and the cloud infrastructure are comprehensively visualized in Fig. 2. This diagram illustrates the integration of the feedback loop from the generative logic back to the physical gate at the edge, ensuring deterministic echo suppression during real-time interaction.

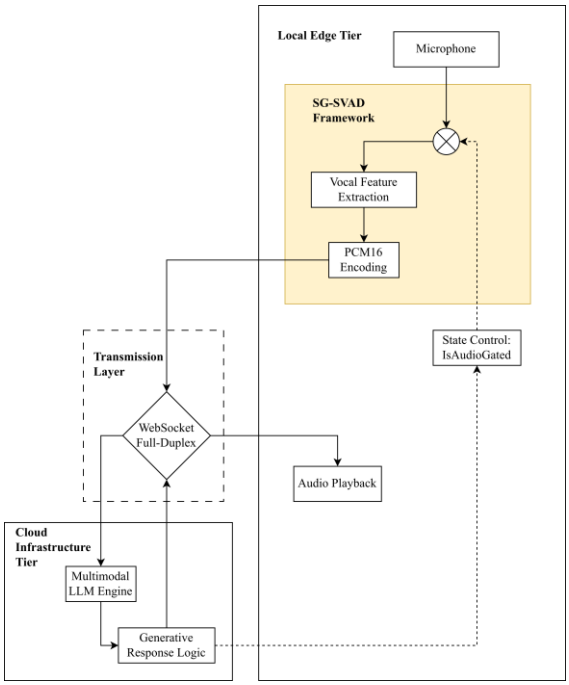

Fig. 2. Hybrid edge-cloud interaction architecture showcasing the full-duplex telemetry and state-control loop

IV. Results And Discussion

A. Experimental Setup

The performance of the proposed SG-SVAD architecture was assessed through extensive testing on four diverse hardware tiers. The specifications of the devices utilized are detailed in Table I. The experiment involved 60 iterations in total (15 iterations per device), conducted in an in-the-wild environment characterized by high acoustic noise in Sumbawa Besar, Indonesia. The network profile recorded via Telkomsel Orbit indicated an average download speed of 76.90 Mbps, upload speed of 10.43 Mbps, and a latency (ping) of 32 ms.

TABLE I
SPECIFICATIONS AND CLASSIFICATION OF EXPERIMENTAL DEVICES

Device Model	Hardware Tier	Chipset/Processor	RAM
Samsung SM-A055F	Entry-Level	MediaTek Helio G85	4GB
Infinix X1102 (Xpad 20)	Entry-Mid	MediaTek Helio G88	8GB
Xiaomi 24117RN76E	Mid-Range	MediaTek Helio G99-Ultra	8GB+
MacBook Pro M1	High-End	Apple M1 (ARM)	8GB

B. Latency and Reliability Performance

The system's performance was measured by pipeline latency (ms) to ensure that the Time-To-First-Audio remained below the human cognitive threshold. As depicted in Fig. 3, the system demonstrates competitive latency performance across all devices. The variability indicated by the error bars reflects the fluctuating network jitter experienced at the testing site, yet the median latency remained within operational limits regardless of the hardware tier. This indicates that the architecture is optimized for real-time responsiveness.

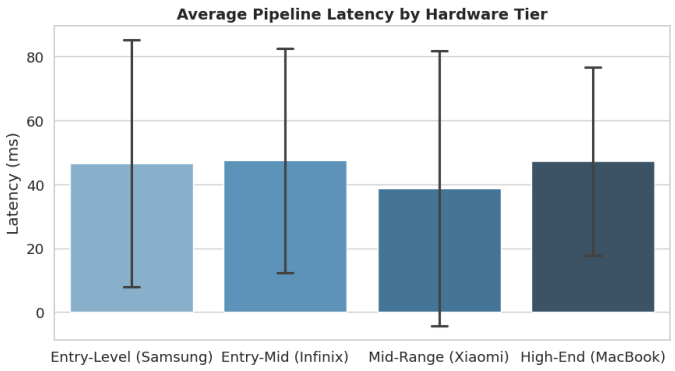

Fig. 3. Average Pipeline Latency by Hardware Tier

C. Computational Efficiency

Computational efficiency was analyzed by measuring RAM consumption during the 3-minute sessions. As illustrated in Fig. 4, the system maintains consistent stability. Average memory usage for Android-based devices remained within the 250–270 MB range. Although the MacBook Pro M1 recorded higher memory usage (706–728 MB) due to the macOS Unified Memory architecture, the system demonstrates exceptional computational power efficiency for multimodal streaming tasks across all platforms. This validates that the state-driven audio gating and deterministic FFT algorithm effectively detect vocal activity without imposing significant CPU overhead.

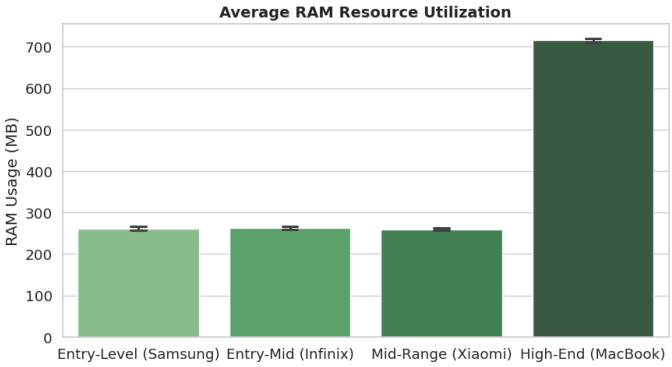

Fig. 4. Average RAM Resource Utilization

D. Comparative Benchmark Analysis

To further validate the technical standing of the SG-SVAD architecture, a comparative analysis was conducted against existing standard architectures found in the literature. As presented in Table II, the proposed method is evaluated against classical baselines and modern neural-based frameworks. While deep learning approaches [11], [15] offer high precision, they demand substantial computational MACs or are limited to simulation environments. SG-SVAD achieves a unique balance by providing the lightweight characteristics of classical methods [3] while ensuring the real-time responsiveness required for modern multimodal interactions on physical hardware.

TABLE II
COMPARATIVE ANALYSIS OF VAD AND STREAMING ARCHITECTURES

Method	Computational Load	Memory Footprint	Latency	Evaluation Env
Wu et al. [3]	Not Reported	< 30 KB	Not Reported	Embedded (PDA)
Yeh et al. [6]	42.98–93.06 M FLOPs	62.2–111.2 k Params	Not Reported	Simulation
Zhang et al. [10]	1.32–8.92 G MACs/s	0.29–0.76 M Params	2.96 ms	Workstation (Laptop)
SG-SVAD (Proposed)	Low (FFT-based)	250–270 MB	< 60 ms	On-device (Edge)

E. Robustness Analysis

System reliability was evaluated based on the success rate of the experimental iterations. As presented in Fig. 5, the proposed architecture achieved a high success rate, reaching up to 80% on higher-tier devices.

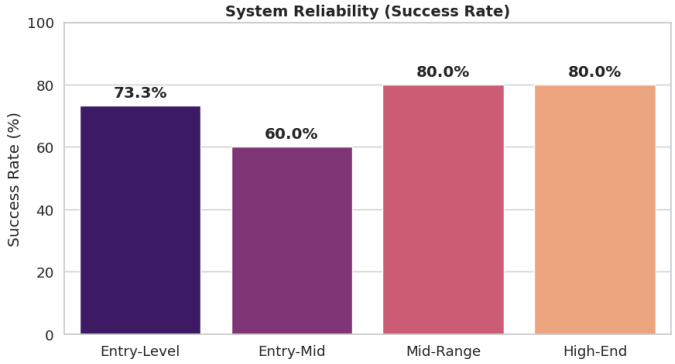

Fig. 5. System Reliability (Success Rate)

Furthermore, the 16 recorded failures during the 60 iterations were analyzed to assess robustness. Failures were categorized into network connectivity failures and inference response failures. Notably, the Entry-Mid tier (Infinix Xpad) exhibited the lowest success rate. The researcher concludes that this is not due to computational inability to execute the VAD algorithm, but rather due to the hardware abstraction layer or communication chipset stability on that specific tablet when handling concurrent streams under unstable network conditions. Conversely, Mid-Range and High-End devices exhibited greater stability. Overall, the achieved success rate under extreme network conditions affirms that the SG-SVAD architecture is sufficiently robust for practical implementation as a language learning assistant in regions with developing digital infrastructure.

CONCLUSION

The real-time multimodal streaming engine, termed SG-SVAD, was developed to address computational and transmission bottlenecks in resource-constrained edge devices, specifically for AI-driven Augmented Reality (AR) interactions. The integration of a zero-parameter spectral Voice Activity Detection (VAD) coupled with a state-driven audio gating mechanism has successfully mitigated the acoustic feedback loops and network overheads traditionally encountered in streaming applications. The empirical evaluation conducted across four distinct hardware tiers confirmed the architecture's efficiency. The system maintained low memory utilization, consistently remaining within the 250–270 MB range on Android devices, while demonstrating robust performance in in-the-wild network conditions. The statistical analysis affirmed that the proposed design is hardware-agnostic, achieving stability regardless of the device's processing tier.

Furthermore, the hybrid edge-cloud streaming approach effectively delegates generative loads via asynchronous WebSocket telemetry, thereby resolving the trade-off between local processing constraints and real-time interaction requirements. The findings indicate that the SG-SVAD architecture provides a reliable foundation for educational applications, particularly for language learning assistants in frontier, outermost, and underdeveloped regions with developing digital infrastructure. Future research will focus on integrating advanced natural language processing models within the cloud-based component to enhance linguistic sophistication, as well as optimizing the VAD algorithm for even lower sampling rates to further conserve computational resources.

REFERENCES

[1] Y. Jani[1] and K. Prajapati[3], "LEVERAGING MULTIMODAL AI IN EDGE COMPUTING FOR REAL-TIME DECISION-MAKING," *International Journal Of Core Engineering & Management*, no. 7, 2023.

[2] I. Froiz-Miguez, P. Fraga-Lamas, and T. M. Fernandez-Carames, "Design, Implementation, and Practical Evaluation of a Voice Recognition Based IoT Home Automation System for Low-Resource Languages and Resource-Constrained Edge IoT Devices: A System for Galician and Mobile Opportunistic Scenarios," *IEEE Access*, vol. 11, pp. 63623–63649, 2023, doi: 10.1109/ACCESS.2023.3286391.

[3] B. Wu, X. Ren, C. Liu, and Y. Zhang, "A Robust, Real-Time Voice Activity Detection Algorithm for Embedded Mobile Devices," 2005.

[4] Y. Huang, W. Qin, Z. Li, and Q. Zhang, "Time-frequency dual-domain attention for acoustic echo cancellation," *Journal of Supercomputing*, vol. 81, no. 5, Apr. 2025, doi: 10.1007/s11227-025-07200-2.

[5] D. Saha, "Towards Automated Interviewing: A Real-Time Conversational Agent Using Speech Recognition and Language Models with Speech Synthesis," Nov. 06, 2025.

doi: 10.36227/techrxiv.176240521.18117168/v1.

[6] D. Egunjobi and O. J. Adeyeye, "Revolutionizing Learning: The Impact of Augmented Reality (AR) And Artificial Intelligence (AI on Education," *International Journal of Research Publication and Reviews*, vol. 5, no. 10, pp. 1157–1170, Oct. 2024, doi: 10.55248/gengpi.5.1024.2734.

[7] H. Mondal and S. Mondal, "Adopting augmented reality and virtual reality in medical education in resourcelimited settings: constraints and the way forward," *Adv. Physiol. Educ.*, vol. 49, no. 2, pp. 503–507, 2025, doi: 10.1152/advan.00027.2025.

[8] T. Wang, J. Guo, B. Zhang, G. Yang, and D. Li, "Deploying AI on Edge: Advancement and Challenges in Edge Intelligence," Jun. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/math13111878.

[9] F. R. Mughal *et al.*, "Adaptive federated learning for resource-constrained IoT devices through edge intelligence and multi-edge clustering," *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-78239-z.

[10] S. Ooko and L. Chaptegai, "Information and Communication Technology for Development (ICT4D) and its Areas of Application," *Mbeya University of Science and Technology Journal of Research and Development*, vol. 6, no. 2, pp. 1–11, Jun. 2025, doi: 10.62277/mjrd2025v6i20002.

[11] Y. T. Yeh, C. C. Chang, and J. W. Hung, "Empirical Analysis of Learning Improvements in Personal Voice Activity Detection Frameworks," *Electronics (Switzerland)*, vol. 14, no. 12, Jun. 2025, doi: 10.3390/electronics14122372.

[12] L. Mu *et al.*, "A Fine-Grained End-to-End Latency Optimization Framework for Wireless Collaborative Inference," *IEEE Internet Things J.*, vol. 11, no. 4, pp. 5840–5853, Feb. 2024, doi: 10.1109/JIOT.2023.3307820.

[13] Y. Ma and A. Nishihara, "Efficient voice activity detection algorithm using long-term spectral flatness measure," 2013. [Online]. Available: http://asmp.eurasipjournals.com/content/2013/1/21

[14] S. Alimi and O. Awodele, "Voice Activity Detection: Fusion of Time and Frequency Domain Features with A SVM Classifier," *Computer Engineering and Intelligent Systems*, May 2022, doi: 10.7176/ceis/13-3-03.

[15] C. Zhang, J. Liu, H. Li, and X. Zhang, "Neural Multi-Channel and Multi-Microphone Acoustic Echo Cancellation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 2181–2192, 2023, doi: 10.1109/TASLP.2023.3282103.

[16] J. Zhou *et al.*, "M2SILENT: Enabling Multi-user Silent Speech Interactions via Multi-directional Speakers in Shared Spaces," in *Conference on Human Factors in Computing Systems - Proceedings* , Association for Computing Machinery, Apr. 2025. doi: 10.1145/3706598.3714174.

[17] M. Hlayel, H. Mahdin, M. Hayajneh, and H. Nurwarsito, "Toward Industry 5.0: A WebSocket–S7 Bridge for Low-Latency, IEC 61588-Compliant Digital Twins in Remote Industrial Automation," *PLoS One*, vol. 21, no. 5 May, May 2026, doi: 10.1371/journal.pone.0342004.

[18] J. Lu, C. Mo, P. Cao, and G. Li, "Design and Implementation of a Real-Time Collaborative Multi-End Interaction Platform Based on WebSocket and WebRTC," in *2025 4th International Symposium on Computer Applications and Information Technology, ISCAIT 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 2181–2186. doi: 10.1109/ISCAIT64916.2025.11010474.

[19] N. Garg, A. Kaur, F. Ahmad, and R. Dutta, "Augmenting Education: The Transformative Power of AR, AI, and Emerging Technologies," 2025, *John Wiley and Sons Inc*. doi: 10.1155/hbe2/5681184.

[20] J. Divya Mary, "Bridging the Gap between Theory and Practice in Rural Education through Digital Twin Technology," 2025.

[21] N. F. Abdullah *et al.*, "Micro-credential for Industry 4.0: engineering safety and reliability with integrated machine learning and soft skills," *Cogent Education*, vol. 13, no. 1, 2026, doi: 10.1080/2331186X.2025.2598923.

[22] M. H. Bijoy *et al.*, "Towards Sustainable Agriculture: A Novel Approach for Rice Leaf Disease Detection Using dCNN and Enhanced Dataset," *IEEE Access*, vol. 12, pp. 34174–34191, 2024, doi: 10.1109/ACCESS.2024.3371511.

[23] Y. Huang, H. Zhu, and C. Chen, "To be the Global Gateways: An Empirical Study on the Infrastructuralization of Large Language Models from the Access Perspective," *Media, Technology and Governance*, May 2026, doi: 10.1177/29777984261433606.