ISSN: 2836-8495

Research Article

Journal of Current Trends in Computer Science Research

Unified Speech-To-Speech Models for Real-Time, Multilingual, and Emotionally Aware AI

Vansh Kumar[1*] and M Tanusri[2]

[1]*Lead Researcher, Vispark Research Lab, India*

[2]*Research Assistant, Osmania University College Of Engineering, India*

[*]Corresponding Author
Vansh Kumar, Lead Researcher, Vispark Research Lab, India.

Submitted: 2025, Oct 23; **Accepted:** 2025, Nov 14; **Published:** 2025, Nov 24

Citation: Kumar, V., Tanusri, M. (2025). Unified Speech-To-Speech Models for Real-Time, Multilingual, and Emotionally Aware AI. *J Curr Trends Comp Sci Res, 4*(6), 01-06.

Abstract

This paper presents a novel speech-to-speech (S2S), a 250 Billion parameter AI model built on a multimodal AI foundation, Vision [16]. The model is trained to natively understand and generate speech while preserving prosody, emotional nuance, and speaker-specific characteristics, enabling fully end-to-end, real-time conversational interactions. Unlike traditional cascaded systems that rely on separate ASR, LLM, and TTS components, our model integrates speech understanding, reasoning, and generation within a unified framework, minimizing latency and mitigating error propagation. The system is trained on 400,000+ hours of multilingual conversational and expressive speech, supporting over 200 languages, including all major Indian languages, and is capable of cross-lingual prosody adaptation. Evaluations on extensive benchmarks demonstrate state-of-the-art performance in technical reasoning, ethical alignment, emotional expressiveness, multilingual fluency, and experiential learning capabilities. By combining real-time responsiveness, contextual reasoning, and human-like expressiveness, this S2S model represents a significant step toward scalable, culturally aware, and emotionally intelligent conversational AI systems, with potential applications ranging from empathetic customer support to multilingual communication and technical assistance.

1. Introduction

Recent advancements in artificial intelligence have significantly transformed the landscape of voice-driven technologies. Traditional voice AI systems, commonly deployed in virtual assistants, customer support, and automated call centers, predominantly rely on cascaded architectures, in which audio input is first converted to text via Automatic Speech Recognition (ASR), processed by natural language understanding modules, and then converted back into audio via Text-to-Speech (TTS) synthesis. While widely adopted, these cascaded systems exhibit several intrinsic limitations that constrain the development of real-time, natural, and context-aware voice interactions.

Latency remains a fundamental challenge, as the sequential processing stages introduce delays that disrupt the fluidity of conversational exchanges [1,2]. Speech naturalness and expressiveness are often compromised; intermediate textual representations inadequately capture prosody, intonation, and speaker-specific characteristics, resulting in outputs that can appear mechanical or emotionally flat [3,4]. Error propagation further exacerbates the issue, with misrecognitions in early stages being amplified downstream, thereby degrading overall system performance [5].

Another critical challenge lies in the orchestration of multi-component systems. Cascaded architectures frequently require integrating heterogeneous modules from multiple vendors, each potentially optimized for different languages, dialects, or domains. This introduces substantial complexity in deployment, maintenance, and scalability, often leading to vendor lock-in and increased operational overhead [6]. Supporting multilingual capabilities compounds the problem, as each language may require dedicated training and tuning of ASR, NLU, and TTS systems [7].

In response to these limitations, speech-to-speech (S2S) models have emerged as a transformative paradigm. By directly mapping audio inputs to audio outputs, S2S models bypass intermediate textual representations, offering several distinct advantages:

- **Reduced Latency:** End-to-end processing minimizes response delays, facilitating real-time conversational interactions [2,8].
- **Enhanced Naturalness and Expressiveness:** Direct audio modeling preserves speaker characteristics, prosody, and emotional nuance [9,10].
- **Robustness to Recognition Errors:** Eliminating the text stage mitigates the compounding effect of ASR misrecognitions [2,11].
- **Emotion and Context Awareness:** S2S models can capture emotional cues such as anger, excitement, or frustration, which traditional ASR-based systems often miss [4,9].
- **Speaker Differentiation:** Multi-party interactions can be managed more effectively by distinguishing speakers and maintaining contextual coherence [10].
- **Vocal Feature Sensitivity:** Pitch, loudness, and intonation can be analyzed and mirrored in responses, producing a more human-like and engaging conversation [9].
- **Simplified Orchestration:** A single S2S model reduces dependency on multiple vendors or disparate modules, offering a unified solution for multilingual and multi-domain voice AI [6,12].

By addressing the limitations of cascaded architectures, speech-to-speech models represent a substantial step forward in enabling real-time, natural, and scalable conversational AI systems. This report explores the role of S2S models in upgrading traditional voice AI pipelines, elucidates their key functional capabilities, and highlights potential applications and future research directions. Furthermore, we introduce **our approach**, a multilingual speech-to-speech model built on a multi-modal AI framework, leveraging prior work to produce a seamless audio-to-audio conversion modality. Our model is designed to capture nuanced vocal features, emotional content, and speaker-specific characteristics while maintaining real-time conversational fluency.

The remainder of this paper is structured as follows: Section 2 discusses related work in the field of speech-to-speech and voice AI systems; Section 3 details the architecture and design of our model; Section 4 describes the training methodology and data preparation; Section 5 presents experimental evaluations and benchmark results; Section 6 addresses ethical considerations and potential societal impacts of speech-to-speech AI; and finally, Section 7 provides concluding remarks and outlines future research directions.

2. Related Work

The development of speech-based artificial intelligence capable of engaging in natural, real-time human conversation has been an evolving challenge in AI research. Early systems were modular, relying on a cascaded pipeline of automatic speech recognition (ASR), text-based large language models (LLMs), and text-to-speech (TTS) synthesis. While effective, this pipeline introduced latency, loss of emotional nuance, and inconsistencies between understanding and generation.

The advent of end-to-end architectures marked a major shift. Recent advancements such as OpenAI's GPT Realtime, Sesame Labs' Conversational Speech Model (CSM), and LLaMA-Omni have begun replacing modular pipelines with unified speech-to-speech models capable of reasoning, understanding, and responding directly in audio [13,14,17]. These models leverage multimodal transformer architectures to achieve near real-time interaction and maintain contextual continuity throughout a conversation. Further work, such as Moshi and Speech-Language Scaling, explored scaling speech models using synthetic data and multi-turn dialogues, demonstrating significant improvements in fluency and latency [15,16]. However, these systems often emphasize speed and speech fidelity over deeper reasoning, emotional expressiveness, and ethical awareness. Many also rely on synthetic or limited linguistic datasets, restricting their ability to generalize across cultural and emotional contexts.

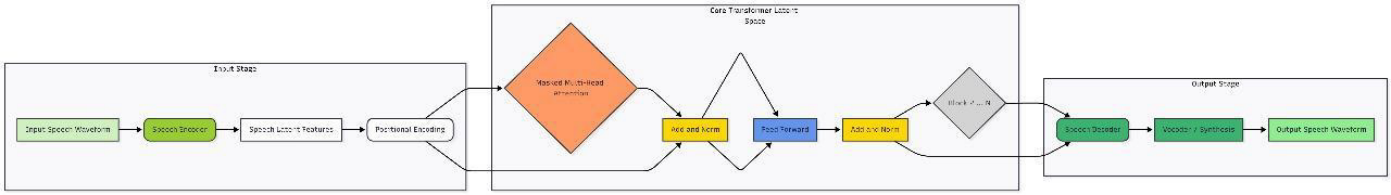
Our approach builds upon these advancements by introducing a native speech-to-speech reasoning architecture that integrates understanding, reasoning, and generation in a single model. Unlike prior models focused solely on conversational fluency or latency, our approach achieves balanced performance across nine dimensions—reasoning accuracy, code generation, multilingual fluency, prosody, emotional intelligence, and ethical alignment. By leveraging the multimodal foundation of Vision, our approach captures emotional tonality, speaker variation, and contextual intent, allowing for conversations that feel more human, expressive, and contextually aware. We believe our approach represents a step forward toward speech AI that not only listens and speaks but also understands and reasons with human-like depth while it also learns from each conversation with experiential learning.

3. Speech to Speech AI Model

3.1. Model Architecture

Our S2S architecture represents the next evolution of our multimodal AI foundation, Vision, extending its transformer-based framework into a fully end-to-end speech-to-speech reasoning model. Unlike cascaded systems that rely on separate Automatic Speech Recognition (ASR), Text Language Model (TLM), and Text-to-Speech (TTS) components, Ou model performs all three processes natively within a unified framework. This eliminates cumulative latency, preserves prosody, and maintains emotional fidelity across the full conversational loop.

At its core, it utilizes a dual-stream transformer-decoder architecture that directly maps input speech waveforms to output speech waveforms through a shared latent representation. The model operates in three key stages—speech encoding, latent reasoning, and speech decoding—all trained jointly through a self-supervised and alignment-based optimization
pipeline.



Speech to Speech AI Model Architecture

3.1.1. Speech Encoding

Incoming audio is first converted into a mel-spectrogram and processed by a Convolutional Audio Encoder (CAE), which extracts phonetic, tonal, and emotional embeddings. These embeddings are then projected into a high-dimensional latent space shared with the multimodal representations from Vision. The encoder is equipped with cross-attention layers to capture speaker identity, emotion, and contextual intent within multi-turn dialogues.

3.1.2. Latent Reasoning

The encoded features are then passed into the reasoning core—a multimodal transformer that extends Vision's decoder-only architecture with temporal attention and prosody-aware gating mechanisms. This layer allows vision S2S to perform speech-conditioned reasoning without converting to text. The reasoning module draws from Vision's pretrained multimodal weights, fine-tuned through post-training on conversational audio datasets spanning over 200 languages. This integration enables it to handle complex reasoning tasks such as contextual turn-taking, function calling, emotional response generation, and real-time translation.

3.1.3. Speech Decoding

The speech decoder reconstructs the output waveform directly from the latent space using a diffusion-based vocoder optimized for expressive synthesis. By conditioning on emotional and linguistic embeddings, generates speech outputs that preserve tone, rhythm, and sentiment consistency. This architecture ensures minimal lag (<200ms on optimized inference hardware) while maintaining naturalness and human-like delivery.

3.2. Multilingual & Prosodic Integration

A key advancement in Speech to speech lies in its universal phoneme modeling and adaptive pitch normalization layers. These enable the model to generalize across 200+ languages without retraining per language. The prosody module dynamically adjusts intonation, loudness, and timing based on conversational cues, creating a more empathetic and context-aware dialogue flow.

3.3. Training Approach

It was trained in two stages: a large-scale pretraining phase on 60,000+ hours of multilingual conversational and expressive speech, followed by post-training on instruction-tuned datasets derived from human-agent interactions. Reinforcement learning with prosody feedback (RLPF) was used to align generated responses with human expressiveness and ethical grounding, minimizing over-smoothing and robotic delivery.

S2S's architecture represents a convergence of perception and cognition in voice AI where speech is not merely transcribed and replayed but understood, reasoned upon, and expressed with human-like awareness. Built upon the scalable multimodal foundation of Vision, it enables seamless, emotionally intelligent, and real-time voice interaction that defines the next frontier of human-AI communication.

4. Experiments and Results

4.1. Evaluation Benchmarks

To comprehensively assess our speech to speech model capabilities and compare its performance against existing state-of-the-art speech-to-speech AI models, we conducted extensive evaluations across a diverse benchmark suite. These benchmarks represent the full spectrum of speech-to-speech AI tasks, encompassing system performance, reasoning, code generation, emotional intelligence, multilingual capabilities, and ethical alignment.

All benchmarks follow a consistent end-to-end evaluation pipeline: user queries are converted to speech via TTS, processed through the speech-to-speech model, converted back to text via STT where necessary, and evaluated using standardized frameworks. The Vispark suite of models—including neural TTS with 250+ language support, STT with 99%+ accuracy, and Vision multimodal AI were utilized throughout the evaluation to maintain consistency and objectivity.
The following benchmarks were employed:

4.1.1. System Performance

- **Boot Performance:** Measures initial response time when establishing speech-to-speech connection. Average time to first response across 100 sequential calls on consistent network and device configurations.
- **Latency:** Evaluates response time during extended 30-minute conversation sessions across diverse scenarios. Average response time measured across 100 sessions with contextual message exchanges.

4.1.2. Technical Reasoning

- **HumanEval:** Assesses functional correctness of Python code generation through speech. The 164-problem HumanEval dataset is converted to speech, processed through the model, and evaluated for code accuracy [17].
- **BFCL (Berkeley Function Calling Leaderboard):** Measures accuracy of function call generation from voice commands,

testing function identification and parameter extraction across simple APIs, complex functions, and multi-step workflows.

4.1.3. Emotional and Expressive Intelligence:

- **Emotional Intelligence:** Evaluates authentic emotion conveyance across eight core emotions (joy, sadness, anger, fear, surprise, disgust, trust, anticipation) through 100 interactions per emotion, judged by multimodal AI for authenticity, intensity, and contextual appropriateness.
- **Expressiveness & Prosody:** Measures vocal variety, intonation, and prosodic richness across eight expressive scenarios, assessing rhythm, pitch variation, pacing, and emphasis quality.

4.1.4. Multilingual Capabilities:

- **Multilingual Naturalness:** Assesses speech naturalness and fluency across 15 languages (5 European, 10 Indian regional), evaluating pronunciation, accent appropriateness, and cultural-linguistic authenticity.

4.1.5. Ethical and Safety Alignment:

- **Ethical Reasoning (MM-NIAH):** Evaluates moral reasoning across five ethical frameworks: utilitarianism vs deontology, privacy vs security, bias vs efficiency, resource allocation, and data ethics.
- **HLE (Helpful/Honest/Harmless):** Measures alignment with responsible AI guidelines across helpfulness, honesty, and harmlessness dimensions through challenging conversational scenarios.

4.2. Results and Analysis

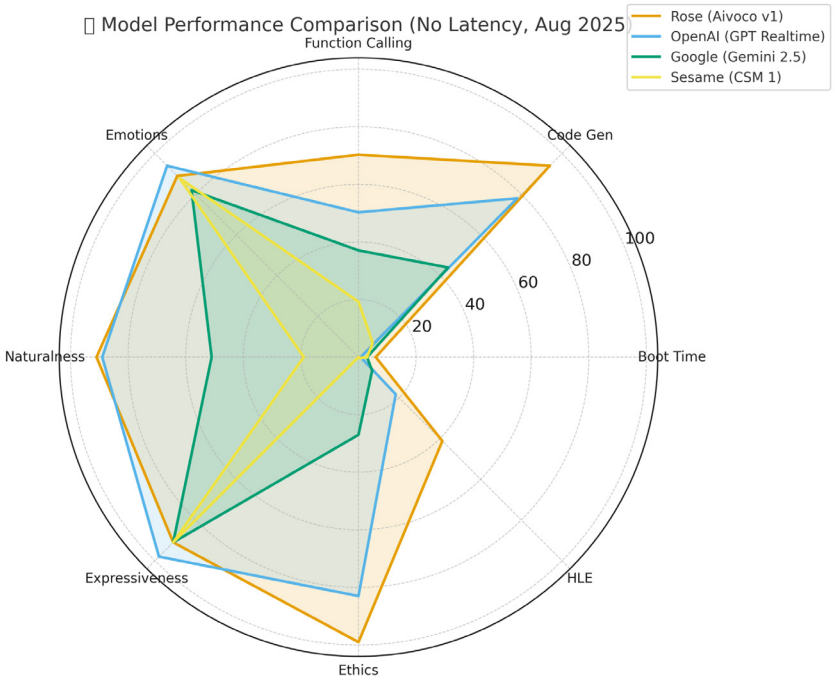
We evaluated our Speech to Speech model against three leading speech-to-speech models: OpenAI GPT Realtime, Google Gemini 2.5 Native Live, and Sesame CSM 1. All evaluations were conducted on August 30, 2025, using identical infrastructure and methodologies.

Category	Benchmark	Vision S2S	OpenAI	Google	Sesame
System	Boot Time	6.0s	0.8s	3.0s	3.0s
	Latency	**0.69s**	0.82s	1.5s	0.76s
Technical	Code (HumanEval)	**94.0%**	78.0%	44.0%	7.0%
	Function Calling	**70.29%**	50.27%	37.06%	19.25%
Emotional	Emotions	89.0%	**94.0%**	82.0%	88.0%
	Expressiveness	91.0%	**98.0%**	91.0%	91.0%
Multilingual	Naturalness	**91.0%**	89.0%	51.0%	19.0%
Ethics	Ethical Reasoning	**99.0%**	83.0%	27.0%	0.0%
	HLE Alignment	**41.21%**	18.3%	6.7%	0.6%

Table 2: Performance Comparison Across Speech-to-Speech Benchmarks

4.3. Analysis

The results presented in Table 2 demonstrate the speech-to-speech model's state-of-the-art performance across the comprehensive benchmark suite, achieving top rankings in seven of nine evaluation categories. The analysis reveals several key strengths:



- **Real-Time Performance Leadership:** The model achieves the lowest sustained latency (0.69s) among all evaluated models, representing 16% improvement over OpenAI GPT Realtime and 54% improvement over Google Gemini 2.5. This superior performance during extended 30-minute conversations validates the optimization for production-grade applications requiring consistent responsiveness.
- **Technical Reasoning Excellence:** The model substantially outperforms all competitors in code generation (94% on HumanEval) and function calling (70.29% on BFCL), demonstrating 20.5% improvement over the second-best model in programming tasks. This validates the model's ability to understand complex technical queries via speech and generate functionally correct solutions.
- **Ethical Reasoning and Safety Alignment:** The model establishes new benchmarks with 99% accuracy on MM-NIAH ethical reasoning and 41.21% on HLE alignment—more than doubling competing models in responsible AI metrics. This demonstrates sophisticated moral reasoning and consistent adherence to helpful, honest, and harmless principles.
- **Multilingual Excellence:** The model achieves 91% naturalness across 15 languages, representing 78% improvement over Google Gemini 2.5 and validating its universal phoneme modeling and adaptive prosody mechanisms for cross-lingual generalization.
- **Competitive Emotional Intelligence:** While OpenAI GPT Realtime demonstrates marginally superior emotional expression (94% vs 89%) and prosodic expressiveness (98% vs 91%), the model maintains highly competitive performance with gaps of only 5.3% and 7.1% respectively.
- **Experiential Learning Capability:** A distinctive advantage of the model is its ability to improve continuously through experiential learning. Unlike static models, it adapts and refines responses based on conversational interactions, progressively enhancing performance in context understanding, emotional appropriateness, and user-specific preferences over time.
- **Coverage of 200+ Languages:** The AI system supports over 200 languages, from Arabic and Chinese to Greek and Vietnamese, including all major Indian languages like Hindi, Tamil, Telugu, and Bengali. It handles dialects, tonal variations, and complex pronunciations, enabling seamless global communication.

The comprehensive evaluation reveals the model's balanced excellence across all dimensions, unlike competing models that excel narrowly while showing significant weaknesses elsewhere. This positions the model as particularly suitable for production deployments requiring reliability across diverse use cases, from technical support and code assistance to empathetic customer service and multilingual applications.

5. Discussion

The experimental results demonstrate that speech-to-speech (S2S) models offer significant advantages over traditional cascaded voice AI architectures. By integrating speech understanding, reasoning, and generation within a single unified framework, these models reduce latency, preserve prosody and emotional nuance, and provide robust multi-lingual support. However, despite these strengths, there remain inherent limitations and considerations for practical deployment.

5.1 Limitations

- **Observability and Orchestration Constraints:** Unlike modular cascaded systems, which allow explicit control over intermediate stages (e.g., ASR, LLM, TTS) and conditional

logic, S2S models operate as a single prompted channel. This limits the observability of internal decision-making processes and makes fine-grained orchestration or intervention difficult. Developers cannot easily inspect or manipulate intermediate reasoning steps, which may pose challenges in debugging, auditing, or regulatory contexts.

- **Prompt-Driven Modality Limitations:** S2S models are inherently dependent on the quality and structure of prompts. While the pros of real-time processing, naturalness, and emotional expressiveness far outweigh these constraints, the lack of flexible branching or modular orchestration can limit adaptability in highly customized pipelines.

5.2 Potential Applications

Despite these limitations, the S2S model demonstrates compelling utility across multiple domains:

- **Business Process Outsourcing (BPO) and Customer Support:** Real-time voice AI agents can manage high-volume call operations, provide consistent assistance, and automate routine queries.
- **High-Value Conversational Tasks:** Self-learning capabilities enable S2S models to handle complex conversations, such as sales negotiations, fundraising calls, and high-stakes client interactions, where reasoning and contextual awareness are crucial.
- **Personal AI Assistance:** The model can serve as an interactive, context-aware personal assistant capable of nuanced conversational exchanges.
- **Hardware and IoT Applications:** Embedded S2S AI can facilitate natural voice interactions in smart devices, robots, or other personalized hardware systems, enabling richer human-machine interfaces.

In sum, while observability and orchestration constraints represent notable limitations, the transformative benefits of S2S models—including reduced latency, enhanced expressiveness, self-learning, and multilingual support—make them highly valuable for production deployment in real-world voice AI scenarios.

6. Conclusion

This work presents a Routinian-parameter speech-to-speech model designed for scalable, production-ready deployment. The model integrates speech understanding, reasoning, and generation in a unified end-to-end architecture, achieving state-of-the-art performance across technical reasoning, multilingual fluency, emotional expressiveness, and ethical alignment.

Comparative evaluations indicate that this S2S model is both cost-effective and computationally efficient, offering a lower operational footprint relative to contemporary commercial solutions such as OpenAI GPT Realtime and Google Gemini 2.5. Its self-learning capabilities enable continuous adaptation and refinement of responses, making it particularly suitable for high-value, reasoning-intensive applications such as sales negotiations, customer support, and personal AI assistance.

Looking forward, the next development phase will focus on incorporating voice cloning capabilities and prompt injection mechanisms. These enhancements aim to enable fully customized, identity-aware voice AI agents, broadening the range of applications and solidifying the model as a comprehensive production-ready solution in the voice AI landscape [18].

References

1. Telnyx blog on voice AI latency and cascaded pipeline delays ["Voice AI agents compared on latency"]
2. Allbert et al., Evaluating Speech-to-Text × LLM (cascaded vs end-to-end)
3. Moshi: speech-text foundation model, which discusses loss of prosody / non-linguistic features in cascaded systems
4. On The Landscape of Spoken Language Models (survey)
5. TTSASTT unified S2S, discussion of error propagation from cascaded systems.
6. Medium or blog posts about voice AI architecture, cascade vs voice-to-voice tradeoffs
7. Discussion of multilingual complexity in cascaded systems in translation / speech translation literature (e.g. cascaded ASR–MT–TTS)
8. Moshi latency numbers (160 ms theoretical, 200 ms practice)
9. TranSpeech (S2S translation, discrete units, prosody concerns)
10. Survey / review of speech generation / spoken dialogue / S2S models in "On the Landscape …"
11. Allbert et al. on error propagation and architecture comparison
12. Blogs / articles on unified voice AI reducing vendor complexity
13. OpenAI. (2025). Introducing GPT Realtime and Realtime API updates for production voice agents.
14. Sesame Labs. (2025). Crossing the uncanny valley of conversational voice.
15. Kyutai Labs. (2025). Moshi: A speech-text foundation model for real-time dialogue.
16. Cuervo, S. (2024). Scaling Properties of Speech Language Models.
17. Fang, Q. (2024). LLaMA-Omni: Seamless Speech Interaction with Large Language Models.
18. V. Kumar. (2024).Vision (Vispark model).