International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 560-567
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0076

AI-Powered Voice Calling Agent Platform for Industry-Agnostic Customer Engagement and Support

Rasika Kachore[1], Manisha Mane[2], Aakanksha Patel[3], Shubham Nikumbh[4], Isha Vesvikar[5], Hardik Sheth[6]
[1,2]Assistant Professor, Dept. of AI&DS, Dr. DY Patil Institute of Engineering, Management & Research, Maharashtra, Pune, India
[3,4,5,6]UG Scholar, Dept. of AI&DS, Dr. DY Patil Institute of Engineering, Management & Research, Maharashtra, Pune, India
Emails: *rasika.kachore@dypiemr.ac.in[1], manisha.mane35@dypiemr.ac.in[2], aakankshapatel04@gmail.com[3], shubhamnikumbh19@gmail.com[4], ishavesvikar@gmail.com[5], 2004hardiksheth@gmail.com[6]*

Abstract
Conventional customer interaction and support processes have long been depending heavily on human customer service representatives or cumbersome Interactive Voice Response (IVR) technology. This has led to inefficient timelines for response and poor service quality and scalability in these domains like marketing, medical, banking and government. In order to mitigate these problems and provide a comprehensive solution to customers and service providers alike, this paper proposes Talksy a plug-and-play AI voice calling agent that is completely domain-independent and works on automated inbound and outbound calls based on context-aware AI.The solution would comprise use cases involving Twilio Telephony Services and Groq's large pre-trained models for efficient inference and task completion. In this solution, customer service processes like lead qualification and follow-up reporting would be conducted automatically by Talksy without human intervention. Analysis and studies have shown that this system not only dynamically provides positive results and effectiveness in customer service and human dependency reduction but is also scalable.
Keywords: *AI Voice Agent, Retrieval-Augmented Generation, Large Language Models, Plug-and-Play Voice Assistant, Contextual Intelligence*

1. Introduction
Customer communication systems have traditionally depended on call centers, manual follow-ups and menu-driven IVRs to manage engagement and support. While these systems provide basic automation, they often suffer from rigid workflows, lack of personalization and high operational overhead. As customer expectations shift toward faster, more intelligent and always-available services, organizations face increasing pressure to modernize their interaction pipelines [1]-[3]. In recent years, breakthroughs in speech recognition technology, natural language processing and large language models have made it possible for conversational AI technology to replicate the real-time conversational experience between humans. Notably, these new AI voice assistants are different from traditional chatbots and IVRs in that they utilize contextual problem-solving and execution of external tools to support sophisticated workflows. It has been observed that adopting the concept of AI-driven automation can bring down service costs by as much as 30% and increase customer satisfaction by 20%. However, the majority of the solutions are still domain-specific or heavily customized and incapable of generalizing across domains. Moreover, the constraints imposed by latency and the absence of grounding within the context make them less suitable for real-time voice applications. In this paper, we present the concept of Talksy, an industry-agnostic AI voice agent with the capacity to tackle the challenges through the application of LLM reasoning and the Retrieval-Augmented Generation approach and the use of the low-latency inference engine [4], [5].

2. Implementation
The implementation of this architecture is built to take advantage of the asynchronous capabilities of

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 560-567
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0076

FastAPI as well as utilize real-time telephony webhook interactions for the creation and monitoring of conversations between a customer and an agent. To address the problem of conversation latency, it uses a combination of Llama 3.3 70B Model to power the inference engine and Groq's Language Processing Units that allow inference times in less than one second. The RAG pipeline maintains technical accuracy by utilizing a FAISS vector store and using HuggingFace embeddings (MiniLM-L6-v2) to search for the most relevant next step based upon the custom company's database or manual of troubleshooting steps. Telephony services are managed by Twilio's Voice API which uses TwiML commands for bi-direction communication and the ability for the LLM to self-direct requests to perform functions like scheduling a service or sending emails based upon the user's intent and defined parameters [6]-[8].

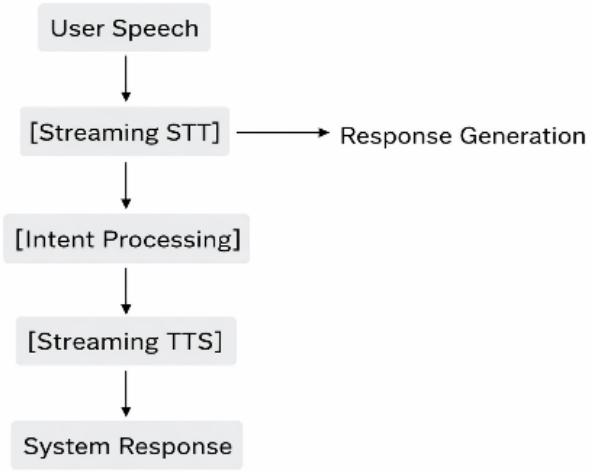

Figure 1 Low-Latency Streaming Architecture for Voice-Based Agents

In figure 1, the streaming and asynchronous nature of the conversational pipeline is illustrated, with speech recognition, intent processing and response generation happening concurrently to ensure there are no latencies in the end-to-end process

2.1. System Architecture

The proposed system adopts a layered, agentic architecture to enable real-time automated voice interactions. Calls (inbound or outbound) are initiated via a telephony service responsible for call setup, audio streaming and capturing speech input. The audio is then processed by a Speech-to-Text (STT) module for real-time transcription and the resulting text is forwarded to a FastAPI-based backend that acts as the core orchestration layer. Within the backend, an agent logic module manages the dialogue flow by interpreting user intent and deciding the next action. Contextual understanding is enhanced through retrieval from a knowledge base containing domain information, conversation history and business rules. Based on this retrieved context, the system determines whether external tools are required, such as scheduling services, notification systems, or CRM integrations, enabling task execution alongside conversation handling. The language model generates the response content, supported by an optimized inference engine to ensure low-latency interaction [9]-[11]. The generated text may be refined through post-processing before being converted into speech via the Text-to-Speech (TTS) module. The synthesized audio is then returned through the telephony layer, allowing continuous conversational flow. The modular design supports independent replacement or optimization of components (telephony, STT/TTS, LLM, vector database), enabling structured evaluation of latency, accuracy and scalability for voice-based customer engagement and marketing applications (Figure 2).

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 560-567
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0076

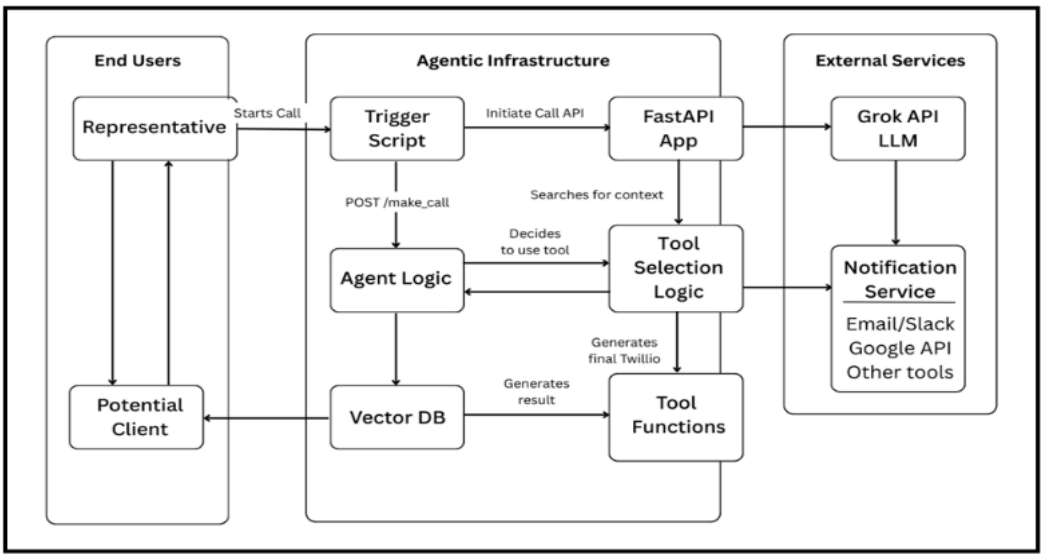

Figure 2 System Architecture

2.2. Tables

The system is made up of different parts that can be easily swapped in and out. This makes it easy to try things and make the system work better.The system uses Twilio for phone calls and speech. It also uses LLaMA 3.3 (70B) to think things through [12]. It does this with the help of Groq.The system uses FAISS to search for things and HuggingFace to understand what things mean. The system is designed to be flexible and to work well. Alternative components such as Telnyx and Plivo for telephony, Deepgram Aura and Whisper for speech-to-text, GPT-4o mini and Mixtral-8×7B for reasoning and ElevenLabs or Cartesia for text-to-speech were evaluated to analyze trade-offs in latency, accuracy and scalability. This modular design enables systematic comparison and efficient deployment of real-time voice-based automated agents for marketing and customer success use cases (Figures 3 and 4).

Component	Current Stack	Alternative 1	Alternative 2
Telephony API	Twilio	Telnyx	Plivo
Speech-To-Text	Twilio Gather	Deepgram Aura	Whisper
LLM Reasoning	Llama 3.3 70B	GPT-4o mini	Mixtral-8x7b
Inference Engine	Groq	Together AI	Ollama
Text-To-Speech	Twilio (Polly)	ElevenLabs	Cartesia
Vector DB	FAISS	Pinecone	ChromaDB
Embedding Model	HuggingFace	OpenAI text-embedding-3	Google Gemini Embeddings
Orchestration	Python/LangChain	Vapi / Pipecat	FastAPI

Figure 3 System Component and Its Alternatives

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 560-567
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0076

3. Results and Discussion

3.1. Results

3.1.1. Latency Evaluation

Latency Component	Mean (ms)	Median (ms)	95th Percentile (ms)
STT Latency	24,800	22,300	52,300
Processing Latency	980	720	2010
TTS Latency	26,100	23,900	54,100
End-to-End Latency	872	690	2070

Figure 4 Component-Wise Latency Summary

The figure shows us the latency of each part in the voice-based pipeline. It uses the middle and 95th-percentile values. Speech-to-text and text-to-speech take a time to process, which is because they have to transcribe and synthesize the entire speech. In contrast the time it takes to process is relatively short which means the system is good at handling what the user wants and coming up with a response. The voice-based conversational pipeline has parts like Speech-to-text and text-to-speech that're important, for the conversation. Despite the high component-level latencies, the end-to-end latency is significantly lower across all statistical measures. This gap highlights the effectiveness of the streaming and asynchronous execution model, where the system generates responses without waiting for complete STT or TTS outputs [13]-[15]. As a result, the system maintains real-time conversational responsiveness even under worst-case conditions.

3.1.2. Plots

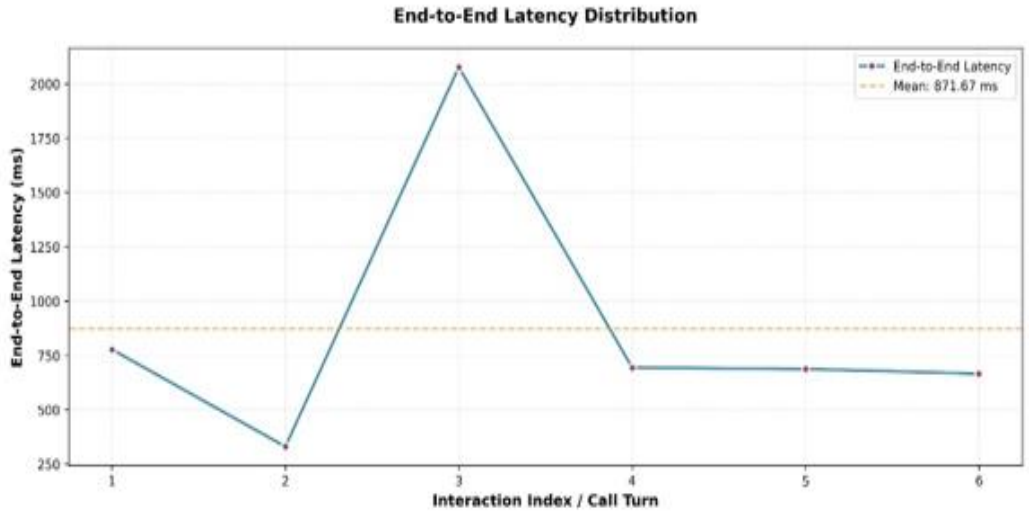

Figure 5 Latency Evaluation Plotting

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 560-567
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0076

Figure 5 shows us the distribution of end-to-end latency across automated call interactions. We see that there are variations in the latency. This happens because of the network conditions and the delays that occur when the system is processing speech. The system is really good at keeping the latency low all the time. When people are talking to the system it usually responds in under one second. This means that the speech recognition and the backend processing and the language model inference are all working together efficiently. The end-to-end latency is still very low, for conversations.

3.1.3. Tables

Figure 6 shows the Comparison of Different LLM Models.

LLM	Typical Accuracy	Streaming Throughput (tokens/sec)
GPT-4o	Very High	~120–150
Claude 3.5 Sonnet	High	~90–110
Gemini 2.5 Pro	High	~70–90
LLaMA 3.1 70B	Medium–High	~35–60
Mistral Large 2	Medium–High	~40–70

Figure 6 Comparison of Different LLM Models

Modern language models like GPT-4o, Claude 3.5 Sonnet and Gemini 2.5 Pro are really good at following instructions and managing conversations. They can have conversations with people one after the other, which is useful for automated phone calls and customer success workflows. These language models, such as GPT-4o, Claude 3.5 Sonnet and Gemini 2.5 Pro, make it possible to have conversations that involve many turns, which is something that language models like GPT-4o, Claude 3.5 Sonnet and Gemini 2.5 Pro are very good at. When we use voice agents how well they work depends on how they're designed. If they can respond quickly and give us a lot of information at once it feels like they are working better. Some things that help are making sure they do not take long to start giving us information and making sure they can give us a lot of information quickly.

3.2. Discussion

This part is about what the experimental results mean for time voice-based automated agents. We do not just repeat the numbers we found. Instead we talk about what the results tell us about how the system's built how well it performs when it has to work fast and how useful it is in the real world, for voice-based automated agents. The new system uses a way of talking that happens at the same time and does not have to wait for each other. This means that the system can hear what you are saying understand what you mean and think of a response, at the same time.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 560-567
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0076

The system does not have to wait for you to finish talking before it starts working. It can start working soon as you start talking. This makes it feel like the system is responding faster. It also makes it feel like a real conversation, where people take turns talking to each other. The system supports this kind of back and forth conversation. The speech recognition and response generation and intent understanding of the system all happen at the time, which is why it can do this. This design choice shows us why the time it takes for things to happen from start to finish stays low. This is true when the time it takes to process speech to text and text to speech is very high. The results tell us that how fast a conversation feels is really about how all the parts work together and overlap. It is not about how long each individual part takes to do its job. The speech-to-text and text-, to-speech components are important. The key thing is the pipeline overlap and how well everything is orchestrated. This is what makes a conversation feel responsive not the time it takes for each component to do its thing. The system works well in real time when we look at how long it takes to respond. This is true when the network and speech are not perfect. Sometimes the system is a little slow to respond. Most of the time it is fast enough for a normal conversation. The system is good, at keeping the conversation going. Latency distribution analysis shows that the system has real-time performance. The latency distribution analysis of the system is important to see how well it works. Latency distribution analysis helps us understand the system. The comparison of different large language models highlights that system-level optimizations play a more important role than model selection alone in voice-based deployments. Techniques such as low time-to-first-token inference, retrieval-augmented generation and constrained tool invocation contribute substantially to response accuracy and reliability.

3.2.1. Use Cases

The proposed framework is designed to be domain-agnostic and extensible, allowing it to be adapted across a wide range of industries. Its modular architecture enables seamless integration with industry-specific knowledge bases, tools and regulatory constraints, making it suitable for deployment in sectors such as healthcare, finance, education, real estate and customer service. By customizing data sources and dialogue logic, the framework can support diverse conversational workflows without requiring fundamental changes to the core system design.

Automated Customer Support: Provides real-time responses to common customer queries and seamlessly escalates complex issues, improving efficiency and response time

Outbound Marketing and Lead Qualification: Conducts automated calls with adaptive dialogue flows, enabling effective lead engagement and reducing call abandonment.

Appointment Scheduling and Reminders: Automates booking, rescheduling, and reminder calls through system integrations, reducing manual effort.

Customer Success, Feedback, and Multilingual Interaction: Supports feedback collection, sentiment-aware conversations, and multilingual communication to enhance user experience across diverse regions.

4. Future Work

The new system is really good at handling voice-based conversations in time. It has a lot of potential for improvement. One thing we can do is make the system work with languages. This means the system will be able to understand and talk to people in languages. We can do this by adding tools that can recognize speeches in many languages understand what people are saying and respond in the same language. This will make the system very useful, for people who speak languages and live in different parts of the world. The system will be able to help people and they will be able to use it more easily. The voice-based automated interactions system will be better because it will work with languages. This is a deal because the voice-based automated interactions system will be able to talk to people in their own language. The voice-based automated interactions system will understand what people are saying and respond in a way that makes sense. This will make the voice-based automated interactions system very popular. People will want to use it. The thing that is really important to work on in the future is making real-time sentiment analysis a part of the system. When the system listens to what people're saying and looks at the words they use it can figure out how they

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 560-567
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0076

are feeling. The system can use this information to change the way it talks to people. It is more understanding and aware of what is going on. This is really useful for things like customer support, where people need to feel like they are being heard and helped. In addition, future efforts will focus on improving scalability and overall system intelligence. Optimizing orchestration pipelines will enable the system to handle high call volumes efficiently while maintaining low latency. Further integration with enterprise tools and analytics platforms will support continuous performance evaluation and dialogue optimization, strengthening the system's capability to deliver personalized and impactful voice-based interactions at scale.

Conclusion

This study confirms that the primary challenge in voice-based automated calling systems—achieving low end-to-end latency despite computationally expensive speech and language processing components—can be effectively addressed through appropriate system design. The results demonstrate that high raw speech-to-text and text-to-speech latencies do not necessarily degrade user-perceived performance when streaming and asynchronous execution are employed. The experimental evaluation shows that the proposed architecture consistently maintains end-to-end latency within real-time conversational thresholds by overlapping speech recognition, intent processing and response generation. The discussion further validates that system-level orchestration, rather than individual model performance alone, plays a decisive role in enabling natural and uninterrupted voice interactions. Overall, the findings confirm that a modular, streaming-based conversational pipeline is well suited for automated marketing and customer success applications. This work provides practical insights into designing scalable voice-based agents and establishes a foundation for future optimization and real-world deployment.

References

[1]. Adikari A, Alahakoon D (2021) Understanding citizens' emotional pulse in a smart city using artificial intelligence. IEEE Trans Ind Inf 17(4):2743–2751. https://doi.org/10.1109/TII.2020.3009277

[2]. Adikari A, Burnett D, Sedera D, de Silva D, Alahakoon D (2021) Value co-creation for open innovation: An evidence-based study of the data driven paradigm of social media using machine learning. Int J Inf Manag Data Insights 1(2):100022

[3]. Adikari A, Gamage G, de Silva D, Mills N, Wong S, Alahakoon D (2021) A self structuring artificial intelligence framework for deep emotions modeling and analysis on the social web. Futur Gener Comput Syst 116:302–315

[4]. Adikari A, Nawaratne R, De Silva D, Ranasinghe S, Alahakoon O, Alahakoon D (2021) Emotions of COVID-19: Content analysis of self-reported information using artificial intelligence. J Med Internet Res 23(4):e27341

[5]. Baevski A, Zhou H, Mohamed A, Auli M (2021) wav2vec 2.0: A framework for self-supervised learning of speech representations. arXiv.org

[6]. Lieskovská E, Jakubec M, Jarina R, Chmulík M (2021) A review on speech emotion recognition using deep learning and attention mechanism. Electronics 10(10):1163 31

[7]. Alahakoon D, Nawaratne R, Xu Y, De Silva D, Sivarajah U, Gupta B (2020) Self-building artificial intelligence and machine learning to empower big data analytics in smart cities. Inform Syst Front. https://doi.org/10.1007/s10796-020-10056-x

[8]. Han K, Yu D, Tashev I (2020) Speech emotion recognition using deep neural network and extreme learning machine. Microsoft Research.

[9]. Hazarika D, Poria S, Mihalcea R, Cambria E, Zimmermann R (2020) ICoN

[10]. K. Akdim and L. V. Casaló, "Perceived value of AI-based recommendations service: the case of voice assistants," Feb. 2023, doi: 10.1007/s11628- 023-00527-x.

[11]. C. Sun, Z. Shi, X. Liu, A. Ghose, X. Li, and F. Xiong, "The effect of voice AI on consumer purchase and search behavior,"

Jan. 2019, doi: 10.2139/ssrn.3480877.

[12]. A. Mari, R. Algesheimer, and N. Outi, "AI-based voice assistants for digital marketing: preparing for voice marketing and commerce," in Digital Marketing, 2021. [Online].Available:https://www.zora.uzh.ch/id/eprint/208002/

[13]. S. Jusoh, "Intelligent conversational agent for online sales," 2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Jun. 2018, doi: 10.1109/ecai.2018.8679045.

[14]. A. Mari, A. Mandelli, and R. Algesheimer, "The Evolution of Marketing in the context of Voice Commerce: A Managerial perspective," science, 2020, pp. 405– 425. doi: 10.1007/978-3-030-50341-3_32.

[15]. S. H. Aldulaimi, M. M. Abdeldayem, B. M. Mowafak, and M. M. Abdulaziz, "Experimental Perspective of Artificial Intelligence Technology in Human Resources Management," pp. 487–511, 2021.