International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 03 | March - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

# AI Calling Agent: Designing Intelligent Voice Systems for Automated Telephone Communication

**Shaikh Saad[1], Motani Saad[2], Siddiqui Zafaryab[3], Shaikh Rafey[4],Ms. Samruddhi Santosh Kamble[5], Ali Karim Sayed[6]**

*[1,2,3,4] Students, Department of AIML, [ANJUMAN-I-ISLAM A. R. KALSEKAR POLYTECHNIC], [NEW PANVEL]*
*[5]Project Guide, Department of AIML, [ANJUMAN-I-ISLAM A. R. KALSEKAR POLYTECHNIC], [NEW PANVEL]*
*[6] HOD, Department of AIML, [ANJUMAN-I-ISLAM A. R. KALSEKAR POLYTECHNIC], [NEW PANVEL*
[1]shaikhsaadxr@gmail.com

Zafaryabsiddqui@gmail.com

Motanisaad14@gmail.com

shaikhrafey@gmail.com

[5]samruddhi456kamble@gmail.com

[6]alikarim.sayed@gmail.com

**Abstract-**

Artificial intelligence is changing the way we communicate and one of the most interesting areas where this is happening is in conversational systems that can talk to people in natural language. ai calling agents are basically systems that can make and receive phone calls on their own and handle the conversation without a human needing to be there. this kind of technology can be very useful in places like customer service healthcare and other sectors where there is a large volume of routine calls that need to be handled every day.

The problem with most systems that exist right now is that they are either too rigid or too fragmented to actually work well in real conversations. the old ivr systems that most companies still use just play a menu and ask the caller to press buttons or say very specific words and if the caller say something slightly different the whole thing break down. this lead to a lot of frustrated callers and still require a lot of human agents to handle the calls that the system cant manage.

Even the newer ai based systems have this problem where they improve one specific technology like speech recognition or language understanding but don't really put everything together in a way that can handle a full phone conversation from beginning to end on its own. because of this companies still depend heavily on human call centers

which is expensive and hard to scale especially when call volume is high.

In this study we try to address this gap by designing and implementing a ai calling agent that combine speech recognition language understanding reasoning and voice synthesis all together in one integrated system. we then test this system on real call scenarios to see how well it actually perform. the goal is to contribute something useful toward building voice communication systems that are more reliable scalable and actually practical to deploy in real environments.

## I. INTRODUCTION

Artificial intelligence has come a long way in terms of making it possible for computers to actually communicate with humans in a natural way. one of the more interesting things that has come out of this progress is the ai calling agent which is basically a system that can pick up or make a phone call and carry out the whole conversation by itself using speech recognition natural language processing and voice synthesis. these kind of agents are starting to be used more and more in areas like customer support healthcare and general service management where organisations have to deal with a very high number of calls on daily basis and cant always have enough human staff to handle all of them.

The ideal situation for any organisation would be that phone calls are handled quickly and accurately and the caller dont have to wait too long or go through a frustrating experience. the idea is that human agents should only need to get involved for the complicated or sensitive cases while all the routine stuff like confirming appointments answering basic questions or giving out information should be handled automatically without needing a real person. but the reality is that the traditional systems like ivr that most organisations currently use are very rigid and the caller often end up frustrated because the system cant understand them properly or force them into a menu that dont match what they actually want. on the other hand maintaining a full human call center is very expensive and difficult to scale up when call volume increase suddenly.

Even though there has been good progress in conversational ai in recent years and speech recognition and dialogue generation has improved a lot many of the existing systems still have trouble managing a full real time conversation properly specially when the context change or the caller say something unexpected. also most of the research that has been done so far tend to focus on improving just one part of the system rather than looking at the whole end to end experience of a complete phone conversation.

So there is clearly a need for a more complete and integrated approach to building ai calling agents that can actually handle real world phone conversations on their own. this study try to address that by looking at how to design and implement such a system using a structured multi layer architecture and then testing it to see how well it work in practice.

## II. LITERATURE REVIEW

Most of the research that exist on ai calling agents is based on earlier work that was done in speech recognition and conversational systems. one of the early important studies was done by alex graves ,abdel-rahman mohamed and geoffrey hinton in 2013 where they showed that deep neural networks can make speech recognition much more accurate than what was possible before. this was a important finding because converting spoken words into text reliably was always a difficult problem and their work helped move things forward a lot. around the same time xuedong huang james baker and raj reddy in 2014 looked at how speech technology had changed over the years and explained how the field was shifting away from systems based on hand written rules toward systems that learn patterns from large amounts of data. but even with all this progress the phone systems that most organisations was actually using at that time like ivr was still very limited. they could only understand a small set of predefined inputs and had no real ability to handle a natural flowing conversation so callers often ended up frustrated.

After that researchers started putting more attention on making the conversational intelligence better. Steve young Milica Gasic Blaise Thomson and Jason Williams in 2013 worked on probabilistic models for managing dialogue which was useful because real conversations are often uncertain and ambiguous and their approach helped systems deal with that better. then Jianfeng Gao Michel galley and lihong li in 2019 looked at how neural networks could be used to improve response generation and make the system more aware of context during a conversation. a really big contribution also came from ashish vaswani noam shazeer niki parmar jakob uszkoreit llion jones aidan gomez lukasz kaiser and illia polosukhin in 2017 when they introduced the transformer architecture. this ended up becoming the foundation that most of the powerful language models today are built on and it made both understanding language and generating responses significantly better than before.

But even with all these contributions there is still a clear gap in the research. most of these studies focus on just one specific component like improving speech recognition or making dialogue management better or enhancing response generation. not many of them actually try to combine all of these pieces into one complete system that can handle a full phone call on its own in real time. there is also not much research that properly look at challenges like keeping the conversation going across multiple turns handling unexpected things the caller might say or making sure the system respond fast enough that the conversation dont feel unnatural.

This is the gap that our study is trying to address. instead of focusing on just one part we are trying to build a complete integrated system that bring together speech processing language understanding reasoning and voice synthesis and then test it in scenarios that actually reflect real world phone conversations..

## III. SYSTEM ARCHITECTURE

The ai calling agent we built use a multi layer architecture that is designed to handle real time voice conversations in a natural way. unlike the old type of automated call systems that just play menus and ask you to press buttons our system actually try to understand what the caller is saying and respond in a sensible way. we connected speech processing language understanding reasoning and response generation all together in one pipeline so each part do its own job but they all work together smoothly to handle the kind of back and forth conversations that happen in real calls.

The first part of the architecture is about taking in what the caller is saying and making sense of it. when someone call the system the voice is captured through the phone interface and sent to the speech recognition module which convert the spoken words into text. this text then go to the natural language understanding part which try to figure out what the caller actually want what information is relevant and what has been said so far in the conversation so the agent dont lose track of the context.

After the system understand what the caller want the dialogue manager decide what the best response would be. it look at the intent that was identified and also consider the flow of the conversation so far before deciding what to say next. this is important because in a real call the caller might ask followup questions or change what they asking and the agent need to handle that without getting confused.

In the last stage the response that was decided get converted into speech using the text to speech module and that audio is sent back to the caller through the call interface. there is also a central integration layer that basically manage all the communication happening between the different modules. this layer make sure data is moving between parts quickly and efficiently so there is no big delay that would make the conversation feel unnatural or broken.

overall this architecture allow the agent to have proper multi turn conversations on its own without a human needing to step in for every little thing.

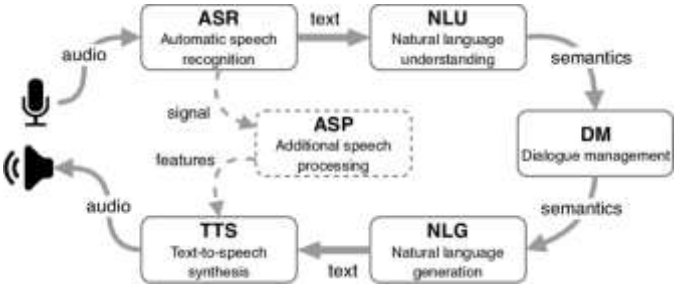


Fig. 1: Block Diagram of AI Calling Agent

The block diagram of the system, illustrated in **Fig. 1**, represents the flow of data across different functional modules of the AI calling agent. The process begins with **voice input**, followed by **speech-to-text conversion**, **intent recognition**, and **dialogue processing**. The system then generates a response, which is converted back into speech and delivered to the user. An integration layer ensures seamless communication between modules and maintains real-time performance. This modular architecture supports scalability, flexibility, and efficient handling of continuous voice interactions.

## .IV. METHODOLOGY

This study use a system based approach to build and test a ai calling agent that can do real time phone conversations on its own. the main idea is to connect all the key parts together like speech recognition natural language processing dialogue handling and voice output in one working pipeline.

The way it work is the system first capture what the caller is saying through the phone line then convert it to text using speech recognition. after that the text go through a language understanding module that try to figure out what the caller actually want and what is the context of the

conversation. based on this the dialogue manager decide what the agent should say next using both some fixed rules and the ai model working together. the reply is then converted back to speech and send to the caller so the conversation can keep going naturally.

To check if the system actually work well we tested it on many different call situations both simulated ones and actual real calls. the kind of tasks we tested include things like asking for information confirming a service or booking appointment. we measure how accurate the responses was how often the call completed successfully and how fast the agent was replying. these measurements help us understand where the system doing good and where it still need more improvement.

## V. IMPLEMENTATION

The implementation of the ai calling agent is basically about taking the architecture we planned and actually making it work as a real system that can do voice calls in real time. the system bring together several ai technologies that all work together to handle speech input figure out what the user want and give back a proper response. first thing we did was setup the telecommunication interface so the agent can receive and also make phone calls on its own.

When a call come in the voice of the caller is picked up and send to the speech recognition module which convert whatever the person is saying into text. this text is then passed to the natural language processing part which try to understand the intent of the caller like what they actually asking for and also keep track of the conversation so far so the agent dont lose context in middle of the call.

Once the system understand what the user want the dialogue manager take over and decide what the agent should reply. this component follow some rules we set and also use the ai reasoning to handle situation that is not covered by rules. it also make sure that if the conversation go on for multiple turns the agent still remember what was said before and dont start repeating or asking same thing again.

After the response is decided it get converted into speech using the text to speech module and that audio is send back to the caller through the same communication interface. there is also a integration layer that sit in the middle of all these modules and make sure they all talk to each other properly with minimum delay so the caller dont feel like the agent is too slow or hanging.

To make it more clear about what each part of the system do we put the main modules and their jobs in Table 1 below.

| Module | Function | Key Role in System |
|---|---|---|
| Telecommunication Interface | Connects the AI system with phone networks | Enables real-time call interaction |
| Speech Recognition | Converts spoken audio into text | Transforms voice input into machine-readable format |
| Natural Language Processing | Analyzes text to detect intent and entities | Enables understanding of user requests |
| Dialogue Management | Determines system responses and tracks conversation context | Maintains coherent multi-turn conversations |
| Text-to-Speech Synthesis | Converts generated responses into spoken output | Produces natural voice replies |
| System Integration Layer | Coordinates communication between modules | Ensures smooth and low-latency processing |

## VI. RESULTS AND DISCUSSION

The ai calling agent was tested through a series of simulated and some real call situations that we designed to match the kind of calls that actually happen in real life. the scenarios we used include things like asking for information confirming appointments and basic customer service type of questions. from what we observed the system was able to handle most of these calls on its own without needing any human to step in which we think show that the overall architecture is working as expected.

In most of the test cases the agent managed to complete the full conversation across multiple turns without losing track of what was being discussed. this suggest that the way we connected the speech recognition the language understanding and the dialogue manager together gave a stable enough base for the agent to work autonomously. the caller didnt have to repeat themselves too much and the

conversation generally felt like it was moving forward properly.

We looked at few important things to measure how good the system is performing. first was response accuracy which basically mean how correctly the agent understood and replied to what the caller said. second was conversation completion rate meaning how many calls actually finished successfully without the agent getting stuck or transferring to human. third was response latency which is how fast the agent reply after the caller stop speaking. the results for all three was generally good as long as the caller was speaking clearly and asking things that was within the kind of conversations the system was trained on. the response time also stayed within limits that felt natural so the caller didnt feel like there was weird pauses.

But we also noticed some problems during testing. sometimes when the caller had a strong accent or there was background noise the system didnt catch what they said properly and either give wrong reply or had to ask them to repeat. also if the caller said something unexpected or went off topic the agent sometimes got confused and the conversation had to be restarted or clarified. these kind of issues is not surprising because same problems have been seen in other conversational ai studies too and it show that even though the technology has improved a lot there is still gap between controlled testing and real messy conversations.

Looking at the bigger picture we think these results show that ai calling agents can genuinely be useful for automating routine phone tasks. companies could use something like this to reduce pressure on their human call center staff for the simple and repetitive calls and only involve real agents for the complicated ones. but to make this work properly the system need to be better at handling different speaking styles noisy environments and unexpected conversation paths. future work should focus on making the speech recognition more robust and giving the dialogue manager better ability to recover when things go off script.

0verall we are happy with the results considering this is a prototype level system. it show that combining multiple ai technologies together in one pipeline can produce something that actually work for real time phone conversations. we hope this work can serve as a useful starting point for others who want to build more advanced voice agent systems in future.

TABLE II

SYSTEM PERFORMANCE METRICS

| Evaluation Metric | Description |
|---|---|
| Response Accuracy | Correct interpretation of user intent |
| Conversation Completion Rate | Successful completion of interaction without human assistance |
| Response Latency | Time taken for system to respond during conversation |
| Conversational Coherence | Ability to maintain logical conversation flow |
| **Evaluation Metric** | **Description** |
| Response Accuracy | Correct interpretation of user intent |
| Conversation Completion Rate | Successful completion of interaction without human assistance |

## VII. CONCLUSION AND FUTURE WORK

In this study we looked at how to build and actually implement a ai calling agent that can handle real phone calls by itself without needing a human to be involved. we connected speech recognition natural language understanding dialogue management and voice synthesis all together in one system and from what we tested it seem to work reasonably well for the kind of tasks we designed it for. the agent was able to handle calls like information requests and service confirmations and in most cases it kept the conversation going properly without losing track of what was being discussed. we think this show that ai has real potential to automate voice based communication specially for the routine and repetitive type of calls that take up a lot of human agent time.

That said we are not claiming the system is perfect or ready for full deployment as it is. there is still some issues that we faced during testing like when callers have different speaking style or accent or when there is noise in background the system Sometime struggle to understand correctly. also if the caller say something unexpected the agent dont always know how to handle it smoothly. these are known challenges in conversational ai and our system is no exception to that.

[1] For future work we think the most important thing is to make the speech recognition more robust so it can handle different voices and noisy conditions better. the dialogue manager also need to be improved so it can deal with more complex or unexpected conversations. it would also be really interesting to test this system on a much larger scale with real users across different languages including hindi and other regional languages to see how well it actually perform outside of controlled conditions. doing that kind of testing would give much better idea about whether this system can be practically used in real call centers or service environments.

Overall we feel this project was a good first step and the results are encouraging even if there is lot more work to be done.

## ACKNOWLEDGMENT

We would like to thank everyone who helped us in completing this research on ai calling agents. there was many people who supported us along the way and without them honestly this project would have been much harder to finish.

First we want to thank our college anjuman-i-islam's kalsekar technical campus new panvel for giving us the facilities and environment to work on this project. special thanks to our project guide ms. fatima khan for always being available whenever we got stuck and for giving us useful feedback that helped us improve our work a lot. her guidance throughout the whole process was really valuable and we appreciate it.

We also want to thank our classmates and friends who sat with us during the difficult parts and gave suggestions even if they was not directly involved in the project. sometimes just talking through a problem with someone help a lot and that happened many times during this work.

we also want to acknowledge the researchers and developers who worked on the open source tools and models that we used in this project. their work made it possible for us to build something meaningful without starting from zero. the wider ai and speech research community has done a lot of hard work and our project is basically standing on top of that.

## REFERENCES

[1] S. Thakur, P. Itankar, P. Gujar, A. K. Sayed, V. Pandey and S. Agrawal, "ER-ADENN:Design and Implementation of EEG-based Emotion Recognition using Adaptive Dropout Enabled Neural Network," 2025 3rd International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, 2025, pp. 320-325, doi: 10.1109/InCACCT65424.2025.11011425. keywords: {Training;Emotion recognition;Adaptation models;Adaptive systems;Accuracy;Sensitivity;Neural networks;Brain modeling;Classification algorithms;Optimization;Emotion recognition;SEED;DEAP;Adaptive dropout enabled network;climbing algorithm},

https://ieeexplore.ieee.org/document/11011425

[2]Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6645–6649.

[3]Huang, X., Baker, J., & Reddy, R. (2014). A historical perspective of speech recognition. *Communications of the ACM*, 57(1), 94–103.

[4]Young, S., Gašić, M., Thomson, B., & Williams, J. D. (2013). POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5), 1160–1179.

[5]Gao, J., Galley, M., & Li, L. (2019). Neural approaches to conversational AI. *Foundations and Trends in Information Retrieval*, 13(2–3), 127–298.

[6]Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 5998–6008.