International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

An Audio: Visual Virtual Personal Assistant

Ateeba Abid[1], Mohammad Tahmeed[2], Harshwardhan Patil[3], Waswi Chinchkhede[4], Ritika Chavhan[5], Dr. G. M Asutkar[6], Mrs. Jaishree Wankhede[7]

[1,3,4,5]Student, Department of Artificial Intelligence and Data Science, Priyadarshini College of Engineering, Nagpur (Maharashtra), India.
[2]Student, Decent College, Nagpur (Maharashtra), India.
[6] Head of Department, Department of Artificial Intelligence and Data Science, Priyadarshini College of Engineering, Nagpur (Maharashtra), India.
[7]Assistant Professor, Department of Artificial Intelligence and Data Science, Priyadarshini College of Engineering, Nagpur (Maharashtra), India.

Abstract
Current traits in smart assistants and smart home automation are currently attracting the interests of customers and researchers. Speech enabled smart virtual assistants (named smart speakers) offer a wide sort of network orientated services and, in some instances, can connect to smart environments, accordingly improving them with new and effective user interfaces. But, such gadgets also reveal new desires and a few weaknesses. Specially, they constitute faceless and blind assistants, not able to reveal a face, and therefore an emotion, and not able to 'see' the person. For this reason, the interaction is impaired and, in a few cases, ineffective. One of the goals of artificial intelligence is the realization of natural talk among humans and machines. In latest years, the dialogue systems, known as interactive conversational systems are the fastest developing vicinity in AI. Many companies have used the dialogue systems technology to set up various sorts of VPAs based on their application areas, including Apple's Siri, Amazon Alexa, etc. To triumph over such troubles, in this project we combine a number of advance techniques. The proposed Assistant is powerful and resource-efficient, interactive and customizable. We use the multi-modal dialogue structures and screen projection which process two or more combined user input modes, which includes speech, image, video, touch, manual gestures, body movements with the intention to design the NextGeneration of VPAs. The new version of VPAs can be used to increase the interaction between humans and machines by the usage of distinctive technologies, consisting of gesture recognition, image/video recognition, speech recognition, dialogue system, conversational information base, and the overall knowledge base. Furthermore, this VPA device can be utilized in different areas such as education, medical assistance, disabilities systems, home automation and security access control.

Keywords: VPA, AI, Voice Recognition, Image Recognition, Home Automation, Gesture Recognition, Motion Detection, Smart Home, Projection, NLP, Python and IOT.

1. Introduction

In recent years, the concept of smart assistants has become widely known and popular. Commercial devices such as Amazon Alexa and Google Home can interact with users using speech recognition and

text-to-speech synthesis, provide multiple network-based services, and communicate with smart home automation systems to provide advanced user interfaces to the system. The popularity of such voice-controlled smart assistants is constantly increasing, thanks to the availability of numerous network services and the increasing number of additional skills and capabilities that can be easily added to smart assistants. However, its potential is still limited by the inability to extract real-time visual information about the user or the environment from video data. This also raises some serious security issues, as most voice-controlled intelligent assistants do not support effective authentication mechanisms but may trigger security-critical actions. Such devices must require facial recognition or other identification mechanisms before accepting voice commands [1]. Current smart assistants can talk to and hear the user, but cannot "see" the user. Additionally, most of the time there is no visual emotional feedback. Because they are blind and faceless to the user, hence, interactions are often impaired and incomplete, and therefore less effective and efficient [2].To solve the above problems, this paper presents an architecture for building expressive and animated graphical characters, as well as visual smart assistants with speech recognition, synthesis and screen projection. The proposed architecture is specifically designed for interfacing with, but not limited to, smart home and home automation platforms. The resulting smart assistant aims to leverage multimodal and nonverbal communication to engage users in highly engaging and effective interactions.

2. Objectives

- **ENHANCED USER EXPERIENCE:** Improve the overall user experience by providing a more intuitive interface. Visual Personal Assistants can make it easier for users to interact with technology and access information or services.
- **COMPANIONSHIP:** The Primary objective of the project is to provide companionship to users. The assistant is designed to interact with users in a friendly and personable manner, offering emotional support and conversation.
- **EMOTIONAL CONNECTION AND ENTERTAINMENT :** The device is designed to create an emotional connection between users and the assistant. Through conversations, gestures, and expressions, the assistant can express empathy and care, helping users feel more connected to it. The assistant can also provide entertainment by singing songs, dancing and engaging in interactive activities that helps users relax and enjoy their time at home.
- **IoT INTEGRATION:** The device can be capable of controlling and interfacing with smart home devices, such as lights, fans, and cameras. It aims to simplify the management of IoT devices and enhance home automation.
- **PERSONALIZATION:** The system can learn from user interactions and adapt to individual preferences. It aims to provide a personalized experience that caters to each user's unique needs and interests. Users can customize the assistant's appearance, outfits, personality and voice.

3. Literature Review

Early research on the effects of embodied virtual agents on human-computer interaction and the "persona effect," i.e the positive effect of the presence of a lifelike character in the interaction environment, dates back more than 20 years [3] . Since then, the original discovery has been confirmed many times in different applications, and the related literature has grown enormously, covering a wide variety of techniques, applications, and approaches [4] [5] [6]. Perhaps one of the most advanced and modern embodied virtual agents is SARA, described in [7]. It has highly precise and complex capabilities for emotional and

expressive human- computer interaction. However, SARA is primarily aimed at emotional interactions by analyzing the user's voice, facial expressions, tone of voice and spoken text. Additionally, SARA is a large and complex system that is difficult to squeeze into inexpensive, small, and resource-limited hardware. Although several other architectures and implementations have been proposed in the literature [8] [9], to the best of our knowledge, an embodied visual and audio-activated virtual agent capable of recognizing users has not been presented. It is open to the public. It can be used on for face detection and can be performed on small, low-cost consumer devices. The lack of facial recognition capabilities in most virtual assistant software can be attributed to the lack of a good, simple, yet effective and accurate approach to facial recognition. In most cases, until a few years ago, facial recognition was either inaccurate or required powerful computing resources for the recognition process or offline registration of the user image [10]. Recent applications of deep neural networks have disrupted this trend and enabled the development of effective, accurate, and lightweight facial recognition techniques [11], which are now also used for smartphone user identification. It’s time to integrate these technologies into your virtual assistant.

4. System Architecture

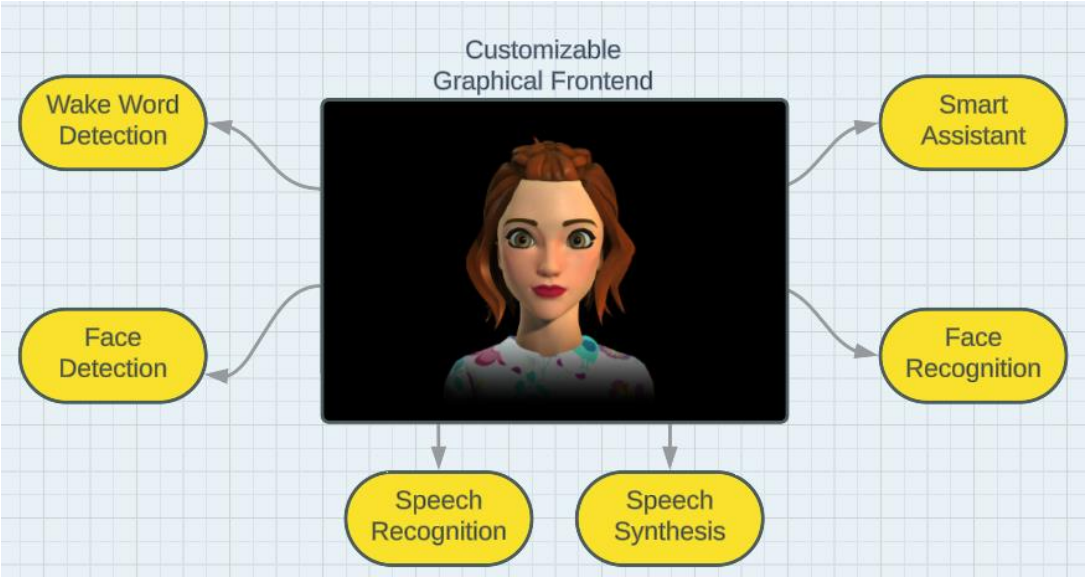

Figure 1: Schematic representation of the modular structure of Hawdraw.

- **Knowledge Base :** The online knowledge base acts as a vast, cloud-hosted wellspring of information. Meanwhile, the local knowledge base stores the data that's more closely tied to our individual modules - things like facial and body datasets for our gesture recognition capabilities, speech dictionaries and dialogue knowledge for the speech engine, video and image datasets for our visual analysis, and even details about our users and system settings.
- **“HAWDRAW” Virtual assistant :**The structure of Hawdraw, as depicted in Figure 1, is a testament to the depth and breadth of its abilities. Each of the six service modules plays a crucial role, whether it's the speech recognition engine that transforms voice commands into text, the gesture module that reads body language, or the face detection system that knows who's interacting with the assistant.
- **The Speech Recognition Model :** It's a sophisticated, end-to-end system that transforms your voice into actionable information. Whether you're issuing a command, asking a question, or simply engaging

in conversation, Hawdraw's speech recognition prowess ensures that your message is heard, understood, and acted upon with precision.

- **Graph Model :** Whether it's recognizing faces, detecting objects, or understanding the nuances of body language, Hawdraw's Graph Model serves as the eyes and brain that transform visual inputs into actionable insights. It's a seamless integration of hardware and software, working in perfect harmony to provide a level of visual understanding that truly sets our virtual assistant apart.

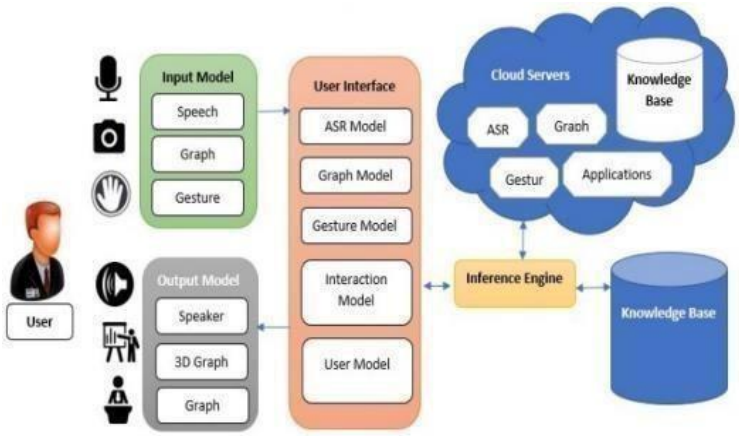

Figure 2 Block Diagram of NextGen VPA

- **Gesture Module:** It's a remarkable feat of technological integration, where hardware and software work in perfect harmony to create an experience that feels almost telepathic. As you interact with Hawdraw, the assistant can intuitively grasp your mood, your needs, and your desired course of action – all through the silent language of your body.
- **Wake Word Detection Module:** It works like this: Hawdraw's Wake Word module is constantly monitoring the audio input, ever-vigilant for the sound of your voice. When it detects a predetermined phrase – similar to "Hey, Alexa" or "Hey, Google" – it springs into action, alerting the Coordination module that you're ready to engage. We can do this by saying “Hello Hawdraw” which is the wake word for our assistant. This name can be personalized as per the user.
- **Face Detection Module:** The Face detection service allows the virtual assistant to detect the presence of a user in front of the device, thus contributing to enable the kind of interactivity that is totally missing in the most common virtual assistants. A Face detection module continuously scans the video input from a connected webcam and, whenever it detects a human face, alerts the Coordination module. When a face is detected, Fox turns its attention to the user facing the camera and greets her. If the identity of the user is available (as the user has been recognized by the Face recognition module), the user is called by name and every change in the identity of the user in front of the camera is signaled by a corresponding utterance (“Hi, how can I help you?”).
- **Interaction Model:** At its core, the Interaction Model is designed to facilitate seamless communication between you, the user, and the intricate network of service modules that make up Hawdraw's capabilities. It's the conduit through which your requests and needs are channeled, ensuring that the appropriate components are engaged to address them.
- **Inference Engine:** This dynamic duo - the Interaction Model and the Inference Engine - form an unbreakable chain, each one complementing the other's strengths. The Interaction Model ensures that

the right information is gathered and routed to the right places, while the Inference Engine applies its advanced analytical capabilities to make sense of it all.

- **Projection Module:** It's a bit like having a digital companion that's able to materialize right before your eyes, yet doesn't obstruct your view of your surroundings. This screen acts as a window into Hawdraw's virtual presence, enabling you to engage with the assistant in a way that feels natural and immersive.
- **User Model:** But this personalization doesn't happen by accident. It's the result of a carefully designed process, where Hawdraw actively engages with its users, asking thoughtful questions and storing the responses in the Knowledge Base. It's a symbiotic relationship, where the more you interact with the assistant, the more it can learn and adapt to provide an experience that feels genuinely customized to you.
- **Input Model:** This model will organize the work of all input devices that the system uses to collect the different data from microphone and Camera. Also, this model includes intelligence algorithms to organize the input information before sending the data to the Interaction Model.
- **Output Model:** This model will receive the final decision from the Interaction Model with an explanation, then it will choose the perfect output device to show the result such data show, speakers or screen based on the result.
- **Libraries used:** Web Browser, Speech Recognition, Pyttsx3, Psutil, Pyjokes, Requests, Datetime, Os. Beautifulsoup, Pygame, etc.

5. Experimental Results

The ongoing development of the virtual assistant included the deployment of a prototype on a Laptop and Monitor to conduct a preliminary evaluation of performance and user experience. The smart assistant architecture featured a modular and customizable graphical interface with animations representing various facial expressions: a vacant expression indicating curiosity, a surprised expression for user recognition and greeting, a silent expression denoting attention to the user, and a chatting expression synchronized with speech.

Figure 3 Picture of the Customizable Prototype

During the test session involving 8 different tasks, users found the experience very positive, expressing interest in continuing the experimentation. To enhance realism and the "persona effect," the assistant's face had subtle movements. Experimentation involved comparing the effects of character animation and face detection/recognition by using a disembodied version of Hawdraw. In this version, without face detection and recognition, users could still interact verbally but weren't recognized individually, and no visual feedback was provided.

TABLE I SUCCESSFULNESS OF TASK IN PERCENTAGE (10 TRIALS FOR EACH TASK WERE PERFORMED)

Task	Successful	Partial Failure	Failure
Open Browser	100%	-	-
Say Wake Word 'Hello'	96.33%	3.67%	-
Ask to tell something about "Hawdraw"	100%	-	-
Ask to Tell a science joke	100%	-	-
Ask to recommend a documentary	100%	-	-
Ask for date and Time in a city	100%	-	-
Find my phone	100%	-	-
Ask for a motivational quote	100%	-	-

Apart from the inability to recognize users and maintain a reliable context-aware interaction, no other significant differences were observed between the sessions. However, analyzing interaction logs revealed users attempting informal interactions with questions like age, origin, and preferences, indicating a desire for conversation. The disembodied agent, lacking visual presence, did not encourage users to view it as a potential conversation partner, leading to both positive and negative implications. The positive aspect was that the embodied agent made interactions more natural and attractive, while the negative aspect highlighted the need for effective natural language processing and a context-aware conversational agent to meet increased user expectations.

TABLE II USER EXPERIENCE QUESTIONNAIRE (REPORTED IN THE FORM OF PERCENTAGE)

Task	Yes	More Yes than No	More No than Yes	No
Do you enjoy the overall experience ?	100%	-	-	-
Did you clearly understand my spoken messages ?	93.33%	6.67%	-	-
Did I promptly catch your commands ?	93.33%	-	6.67%	-
Would you like to have me on your desk at home ?	100%	-	-	-
Would you enjoy my services on a daily basis ?	100%	-	-	-

In summary, the experiment emphasized that visually interacting with the user significantly enhances a virtual agent's appeal. The embodied agent successfully created a more natural interaction, although meeting such expectations requires an effective natural language processing interface and a context-aware conversational agent.

6. Conclusion and Future Work

This project introduces the concept of Next-Generation Virtual Personal Assistants, a new system designed to interact with humans in a more natural and coherent manner. This system utilizes various modes of communication, including speech, graphics, video, and gestures, in both input and output channels. By incorporating technologies like gesture recognition and speech recognition, the system aims to enhance the interaction between users and computers.

One of the key features of this system is its ability to engage in lengthy conversations with users, thanks to its extensive dialogue knowledge base. It can be applied to a variety of tasks, such as providing educational or medical assistance, operating in robotics and vehicles, assisting individuals with disabilities, managing home automation, and controlling security access.

Moreover, this system can serve as a versatile solution for applications such as customer service, training and education, facilitating transactions, online shopping, providing travel information, offering counseling, and functioning as a tutoring system. The final product will be capable of performing tasks such as projection, sensing, speech, and networking

References

1. X. Lei, G. Tu, A. X. Liu, C. Li, and T. Xie, “The insecurity of home digital voice assistants - amazon alexa as a case study,” CoRR, vol. abs/1712.03327, 2017.
2. J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy, “Creating rapport with virtual agents,” in Intelligent Virtual Agents (C. Pelachaud, J.-C. Martin, E. Andre, G. Chollet, K. Karpouzis, and D. Pel ´ e, eds.), (Berlin, ´ Heidelberg), pp. 125–138, Springer Berlin Heidelberg, 2007.
3. J. C. Lester, S. A. Converse, S. E. Kahler, S. T. Barlow, B. A. Stone, and R. S. Bhogal, “The persona effect: Affective impact of animated pedagogical agents,” in Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, CHI ’97, (New York, NY, USA), pp. 359–366, ACM, 1997.
4. Embodied Conversational Agents. Cambridge, MA, USA: MIT Press, 2000.
5. E. Andre and C. Pelachaud, ´ Interacting with Embodied Conversational Agents, pp. 123–149. Boston, MA: Springer US, 2010.
6. B. Weiss, I. Wechsung, C. Kuhnel, and S. M ¨ oller, “Evaluating embodied ¨ conversational agents in multimodal interfaces,” Computational Cognitive Science, vol. 1, p. 6, Aug 2015.
7. Y. Matsuyama, A. Bhardwaj, R. Zhao, O. Romeo, S. Akoju, and J. Cassell, “Socially-aware animated intelligent personal assistant agent,” in Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 224–227, Association for Computational Linguistics, 2016.
8. M. Schroeder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, and M. Wllmer, “Building autonomous sensitive artificial listeners,” IEEE transactions on affective computing, vol. 3, pp. 165–183, 4 2012. eemcs-eprint-22932.
9. B. Martinez and M. F. Valstar, Advances, Challenges, and Opportunities in Automatic Facial Expression Recognition, pp. 63–100. Cham: Springer International Publishing, 2016.
10. F. Battaglia, G. Iannizzotto, and L. Lo Bello, “A person authentication system based on rfid tags and a cascade of face recognition algorithms,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, pp. 1676–1690, Aug 2017.

11. F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823, June 2015