Smart Voice Assistant Using Python

Ravindra Nath[1], Devraj[2*] , Meenakshi Malik[3] and Nikita Singh[4]

[1] Associate Professor, BBAU Central University, Lucknow(U.P.), India.

[2] Principal Scientist(Comp. Appln. & IT), ICAR-IIPR, Kanpur(U.P.), India.

ICAR- Indian Institute of Pulses Research, Kanpur, Uttar Pradesh, India.

[3]Scientist, ICAR- NRC for IPM, New Delhi, India.

[4] Young Professional-II, ICAR-IIPR, Kanpur(U.P.), India.

Abstract

Voice assistants have become integral to human-computer interaction, enabling hands-free operation across devices and domains. It uses all forms of speech technologies, including speech recognition, speech synthesis, a system for creating and analyzing voice data, and voice biometrics, is one of the most sophisticated products in this field. Software that lets you operate your device using voice commands is known as a voice assistant. A modern assistant is capable of far more than just carrying out instructions; they may also carry on a dialogue with the user. A smart assistant's job is to make sure that all of the many technologies that make up a voice assistant function harmoniously together, as they are a complicated innovation. These days, there are a lot of voice assistants available that can improve someone's quality of life. However, as the number of assistant increases annually, selecting one becomes more challenging for the typical user due to the unique qualities of each assistant. More people are using this technology on a daily basis. Games are another industry that is growing quickly. The next natural step, given the increasing number of advances, is to use a voice assistant—at least initially for training purposes.

This paper presents the design and develop of a Smart Voice Assistant(SVA) built primarily using Python and open-source speech and Natural Language Processing(NLP) tools. The SVA integrates state of the art Automatic Speech Recognition(ASR), Tokenization, Support Vector Machine(SVM) algorithm, rule-based technologies, and speech-to-text(STT) components into modular, extensible architecture. This study offers an in-depth analysis of how cutting-edge technologies are reshaping the global economy. As technology continues to evolve at an unprecedented rate, its influence extends to virtually every sector, from manufacturing to services, creating both opportunities and challenges. Our main objective to explore the rise of artificial intelligence, machine learning, automation, and blockchain technology, among others, and their potential to disrupt traditional economic models. We also conclude with the discussion of future directions including enhancement of speech recognition in noisy environment and ability to process multi-steps or context-aware commands.

Keywords: Voice assistant, Support vector machine, Digitization, Python, Speech-to-text, Machine learning, Automatic speech recognition

Introduction

The male gender predominates in the computer sector, and as a result of this blatantly biased representation, many of the products that are made disregard women. Natural language processing is one of the domains that this bias affects. Due to the dearth of women in the field, the identification rate for female voices is lower than that of male sounds, leading to the inquiry and investigation of voice aid selection bias [11,20,21]. These AI-driven applications enable users to interact with devices through natural language, providing a seamless and efficient way to perform tasks hands-free [22,23]. Smart voice assistants, such as Amazon's Alexa, Apple's Siri, and Google Assistant, are increasingly used to control smart home devices, search the internet, play music, set reminders, and more [3]. The rise of voice assistants is driven by advancements in Natural Language Processing (NLP), machine learning, and speech recognition technologies, which have made it possible for machines to understand and respond to human speech with impressive accuracy [1,17]. Speaking and understanding English can be very challenging, particularly if it is not your first language. Not only is it difficult to understand and use English vernacular, but there is also a linguistic barrier in the various dialects that individuals speak. Although this initiative primarily focuses on the English language, a person's dialect can hinder communication in many other languages as well [5]. The history of voice technology traces back to 1962, when IBM introduced shoebox, an early speech recognition system. Innovative solutions emerged in the 1990s when voice assistants could be employed in real-world situations, such as handling phone requests. By 2000, advancements in neural network, cloud computing, and smartphone capabilities led to the development of the first voice assistant with robust user interaction features[2]. Voice assistant such as virtual agents and conversational interfaces have transformed how users interact with computing devices, enabling more natural, hands-free experiences for tasks ranging from information retrieval to home automation. Recent advances in deep learning have significantly improved speech recognition and natural language understanding, while open-source tools and python libraries have democratized development.

The main objective of this study is to design, develop and implement a basic but functional smart voice assistant in python that can carry out tasks like setting alarms, answering general queries, sending messages, and controlling device operations. The voice assistant will employ a combination of NLP techniques and Python programming to recognize user commands and provide contextually relevant responses[8]. Additionally, the study also aims to showcase the potential of smart assistants in personalizing user experiences and enhancing everyday productivity through voice-enabled technology [4]

Voice Assistant System Overview

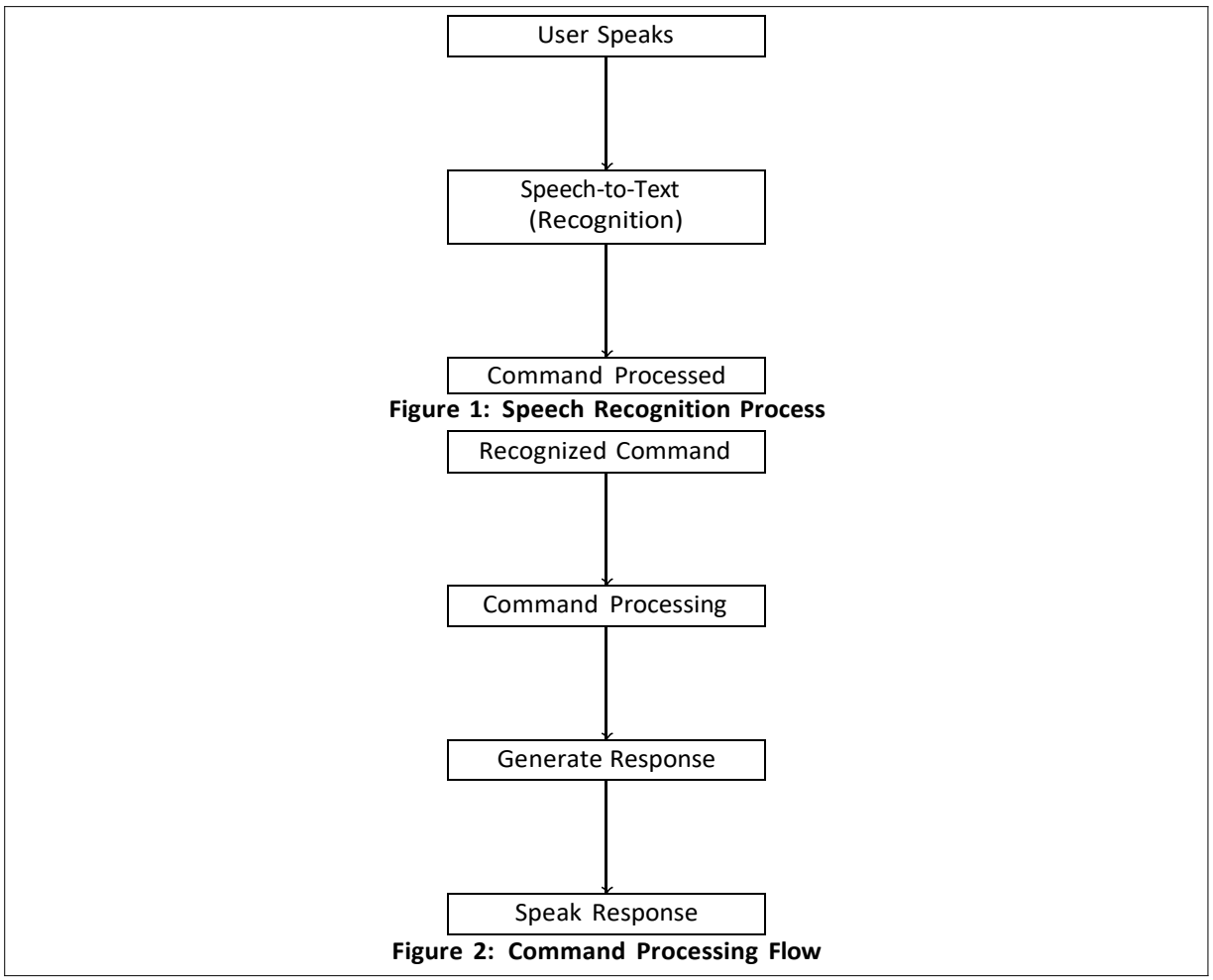

Figure 1: Speech Recognition Process

Figure 2: Command Processing Flow

Figure 3: Voice Assistant System Overview

Review of literature

Voice assistants have a long history ranging from early command-and-control systems to modern conversational agents. Seminal work in speech recognition and signal processing established core techniques for acoustic modeling and decoding. Modern approaches favor neural acoustic models, sequence-to-sequence architectures, and large pre-trained language models for NLU. Commercial systems (e.g., Alexa, Siri, Google Assistant) combine cloud-based ASR/NLU with device-side components. Open-source efforts and toolkits such as Kaldi, Vosk, Mozilla Deep Speech, and more recent transformer-based speech models enable local and research-driven solutions.

Research on assistant design covers dialog management strategies (rule-based, statistical, and hybrid), intent classification and slot filling, multimodal integration, and privacy-preserving on-device inference. In the context of Python, libraries such as Speech Recognition, pyttsx3, and wrapper APIs for cloud services (Google Cloud, AWS, Azure)

simplify development for rapid prototyping. Bapat *et al.*, discussed a significant advancement in speech and statement analysis. They describe how an analog signal is transformed into a spoken signal and subsequently converted into a digital wave used for human-machine communication. This technology, being extensively utilized and infinitely reusable, enables computers to deliver both entertaining and useful services while responding to users' commands in a consistent and reliable manner. The Speech Recognition System (SRS) has a broad range of applications and is a rapidly expanding field. A fundamental model that summarizes the procedure has been developed through research [10]. B. S. Atal and L. R. Rabiner introduced tone analysis as a significant part of speech processing. This approach involves examining speech signal structures to classify them as either pronounced speech, non-verbal expressiveness, or silence. The methodology proposed by Atal and Rabiner is highly effective in speech recognition systems; however, one limitation is the necessity of employing an algorithm within predefined signal sizes under specific recording conditions [9]. T. Schultz and A. Waibel examined the global deployment of speech technology products emphasizing on speech recognition systems. One of the challenges identified in their study is the immobility of addressing new languages. The authors highlight the difficulty of efficiently and methodically implementing vocabulary-intensive systems such as Large Vocabulary Continuous Speech Recognition (LVCSR) systems, especially when expanding to new languages with limited resources [12]. To enhance speech recognition in noisy environments, Faiz, Srivastava, and Khoje presented a virtual voice assistant for smart devices. Their study emphasizes integrating advanced machine learning techniques to enhance task execution and user interaction, providing more accurate responses in diverse environments [15]. Recent research has also scrutinized the real-world implications of speech recognition. Kim *et al.*, discussed the challenges involved in auditory processing and speech recognition in noisy environments. Their research provides methods for improving the robustness of speech recognition systems in challenging acoustic conditions, which is particularly crucial for real-world applications of voice assistants [16]. Robison explored gender biases present in voice assistants. He focused on how the default female voices in many voice recognition systems could perpetuate gender stereotypes. His study calls for a shift toward more inclusive, gender-neutral voice assistant technologies, advocating for both diversity and fairness in the development of such systems [4].

Methodologies

The smart voice assistant follows a modern pipeline with clearly separated stages, enabling experimentation in different components:

Automatic Speech Recognition (ASR Algorithm): It converts spoken language into written text by analyzing audio signals and matching them to corresponding words. It uses an acoustic model to map sounds to phonetic units and a language model to predict word sequences. The system extracts features like MFCCs from the audio and applies techniques such as Hidden Markov Models (HMMs) or deep learning methods (e.g., RNNs, CNNs, and Transformers) for speech recognition [13, 14]. ASR is widely used in applications like voice assistants, transcription services, and automated customer support, enabling voice interaction with technology.

Tokenization: Tokenization involves breaking text into smaller units called ”tokens,” such as words, subwords, or characters. This step is essential in natural language processing (NLP) to help machines understand and analyze text. For example, word-level tokenization splits a sentence into individual words (”I love coding” → [”I”, ”love”, ”coding”]), while subword tokenization might break down complex words into smaller parts (”unhappiness” → [”un”, ”happi”, ”ness”]). Tokenization is used in tasks like text classification, machine translation, and speech recognition, serving as a foundational step in preparing text for further processing by algorithms.

Support Vector Machine (SVM Algorithm): It is a supervised machine learning algorithm used for classification and regression tasks. It identifies by finding the optimal hyperplane that best separates data points into different classes, maximizing the margin between the classes. SVM uses support vectors, the closest data points to the hyperplane, to define this boundary. It can handle non-linear data using the kernel trick, mapping data to higher dimensions where linear separation is possible. Common kernels include linear, polynomial, and Radial Basis Function (RBF). SVMs are widely used in applications like image classification, text classification, and bioinformatics due to their accuracy and robustness.

Statistical Technique: Statistical techniques involves mathematical formulas that are used to analyze data and make predictions applied across fields like research, business, and psychology. These methods enable data-driven insights and forecasting by modeling patterns and relationships within datasets.

Stochastic Technique: A stochastic model is a method for predicting statistical properties of possible outcomes by accounting for random variance in one or more parameters over time. This approach is useful for modeling systems with inherent uncertainties, providing probabilistic forecasts.

Rule-Based Technique: A rule-based technique is a method that uses a set of rules to make decisions or derive conclusions about a problem. Rule-based systems are often made up of if-then rules, and can be designed using expert knowledge or by learning from data offering structured solutions for specific problem domain.

Hybrid Technique: A hybrid technique is a combination of two or more techniques that are used to achieve a better result for a specific problem. Hybrid techniques can be used in a variety of fields, including engineering. These approaches leverage the strengths of multiple techniques to enhance performance.

Interfacing Technique: Interfacing techniques are methods for connecting devices and allowing them to exchange information. The goal of interfacing is to establish a communication path between a computer and its peripheral devices by defining common connection standards and signal meanings.

Results and discussions

The objective of this study is to develop python-based speech-generated virtual assistant comparable to popular voice assistants like Siri, Alexa, or Bixby. The developed smart voice assistant successfully handled a range of basic tasks akin to those performed by mainstream assistants. Upon receiving voice commands, the system executed a limited set of queries, including opening websites, fetching the current time, and telling jokes, with high accuracy and quick response times of 2-3 seconds, ensuring a smooth user experience [24, 27].

- **Opening Websites**: The assistant effectively opened specific websites such as YouTube, Wikipedia, and Google upon receiving commands like ”Open YouTube” or ”Search Wikipedia for Python”. Using predefined functions, the assistant launched the respective URLs in the default web browser.
- **Telling Time**: When prompted with ”What is the time?”, the assistant provided the current time in a clear, spoken response, such as ”The current time is 3:15 PM.”
- **Telling Jokes**: The assistant was also able to respond entertainment-based commands. For example, when the user asked, ”Tell me a joke,” the assistant

responded with a simple joke, such as, ”Why don’t skeletons fight each other? They don’t have the guts.”

- **Search Queries**: For commands like ”Search Google for Python tutorials,” the assistant performed a web search and read out the first result, helping users quickly find relevant information.

While the assistant performed well for these basic tasks, handling complex or multi-step commands remains a challenge. The system’s accuracy and usability make it suitable for simple, everyday tasks, providing a foundation for further enhancements [25, 26].

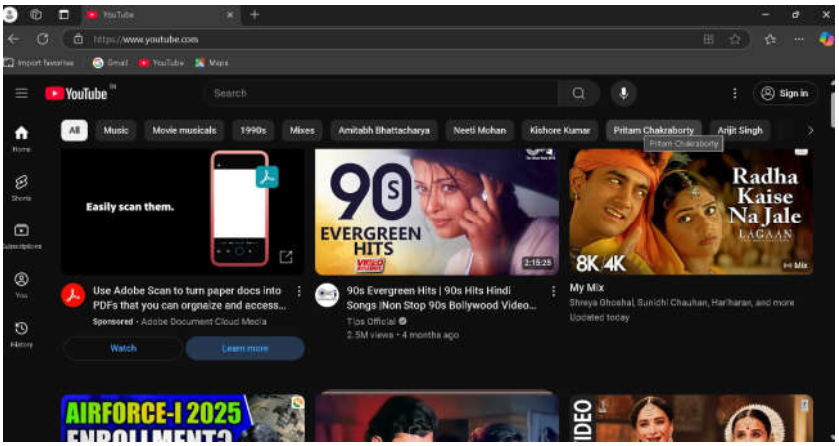

Figure 4: Open Youtube

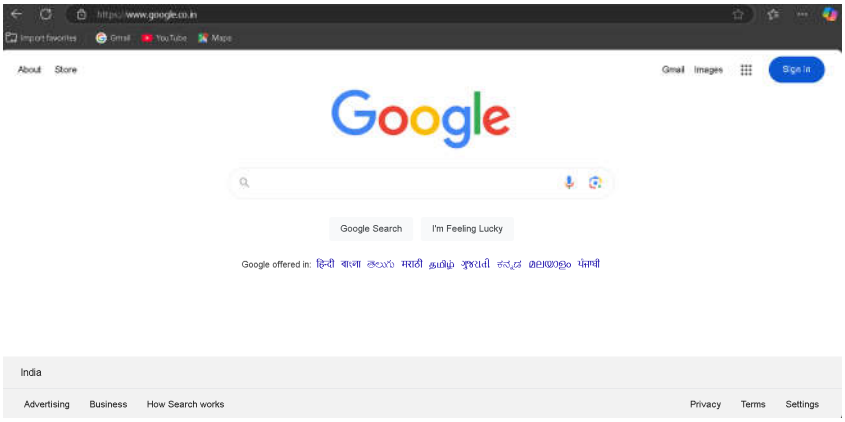

Figure 5: Open Google

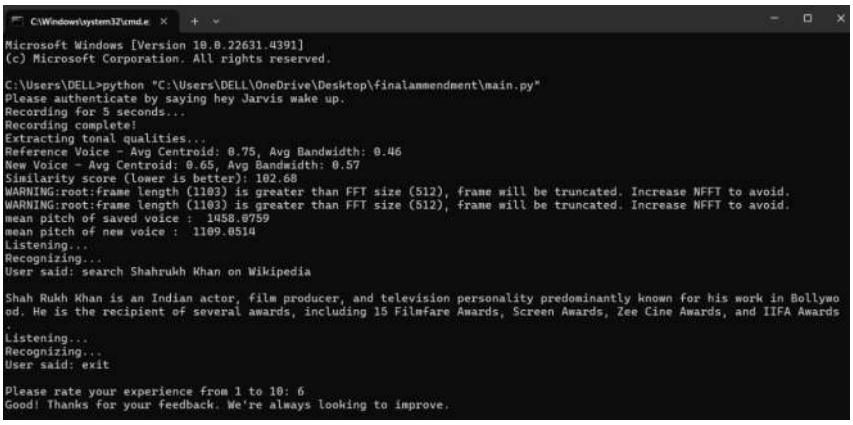

Figure 6: Execution Completed

Tonal quality, or timbre, distinguishes each sound, even if two sounds have the same pitch and volume. When comparing tonal quality, factors like the instrument or sound source, the technique used to play it, and the acoustics of the environment are important. Different instruments produce distinct tonal qualities because of how they're built and how they create sound [18]. For example, violins create sound through vibrating strings, while wind instruments rely on air columns. To understand tonal quality, we can look at both technical measurements, like frequency analysis, and descriptions like "warm" or "bright." This helps us understand how we perceive sound in different settings.

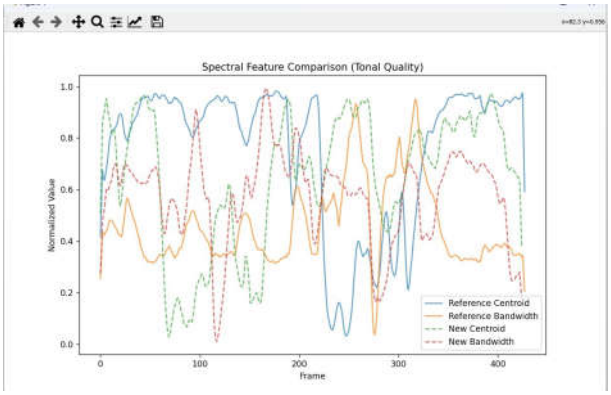

Figure 7: Tonal quality Comparison

The analysis highlights the challenges encountered during the development of this platform and how it distinguishes itself from other existing solutions [6]. We have faced so many challenges during developing system as one of the authentication process and another is the noise cancellation. In terms of accuracy, the system performed well in quiet environments, with correct recognition of voice commands such as "Open Google" or "What is the time?"[7].

However, accuracy decreased in noisy settings, highlighting the need for

improved noise filtering. The assistant's response time was typically within 5-10 seconds, ensuring a smooth user experience for simple tasks. While the system was reliable for its predefined commands, it struggled with more complex or multi-step requests. After executing or exiting commands, we ask for feedback, and based on this feedback, we rate the program on a scale of 1 to 10. User feedback was generally positive, with satisfaction in handling basic queries. However, the assistant's functionality was limited to a small set of commands, which restricted its versatility. Users also mentioned that the assistant could be improved by adding more features, such as the ability to recognize different voices or handle more intricate requests, which would make it more useful in a variety of situations.

Conclusions

The smart voice assistant developed using Python proved to be effective for performing simple tasks such as opening websites, telling the time, and sharing jokes. It responded quickly and accurately to basic commands in quiet settings, providing a user-friendly experience. However, the assistant faced challenges when handling more complex or multi-step commands, indicating that it is still limited in its capabilities. Additionally, background noise impacted its speech recognition accuracy, showing a need for improvements in noise handling.

Despite these constraints, the assistant demonstrates the potential for more advanced applications. By expanding the range of commands it can understand and refining its speech recognition, the assistant could become more versatile and reliable. Future developments could focus on improving its ability to handle more complex tasks, enhancing natural language understanding, and ensuring better performance in various environments. Overall, while the assistant is a solid starting point, further work is required to make it a more practical and capable tool for everyday use.

Future directions

To enhance the smart voice assistant's effectiveness, key improvements include expanding the command set, improving speech recognition in diverse environments, enabling complex task execution, and advancing natural language understanding. The tonal quality of the voice, encompassing pitch, warmth, and clarity, was analyzed and compared using a graph referenced in the results section.

Specific areas for improvement include:

1. **Expanded Command Set**: The assistant currently supports basic tasks like opening websites and telling the time. Adding advanced features, such as setting reminders, controlling smart home devices, or sending messages, would enhance its practicality for daily use.
2. **Improved Speech Recognition in Noisy Environments**: While effective in quiet settings, the assistant struggles with background noise. Implementing advanced noise-canceling algorithms or leveraging cloud-based recognition could improve accuracy in real-world conditions.
3. **Multi-Step and Context-Aware Command Processing**: The assistant is limited to single, direct commands. Enabling it to handle sequential tasks, such as "Set an alarm for 7 AM and play relaxing music afterward," would increase its utility.
4. **Enhanced Natural Language Processing (NLP)**: Improving NLP capabilities would allow the assistant to understand complex or conversational language, enabling it to process a broader range of queries.

By integrating these enhancements, the assistant could evolve into a more powerful, flexible, and user-friendly tool, delivering a richer and more efficient experience for users.

References

[1] Caliscan, A. and Griffin, M. (2021). Bias in natural language processing. *Annual Review of Linguistics, 7*, 523-546.

[2] Soofastaei, A. (2021). Introductory chapter: Virtual assistants. In Book Chapter, A. Soofastaei (Ed.), *Virtual assistant,* 1-10. https://doi.org/10.5772/intechopen.95329.

[3] Berdasco, A., Lopez, G., Diaz, I., Quesada, L. and Guerrero, L. A. (2019). User experience comparison of intelligent personal assistants: Alexa, Google Assistant, Siri and Cortana. *Proceedings of the 5th International Conference on Interactive Mobile Communication, Technologies and Learning (IMCL)*, 179-187. https://doi.org/10.3390/proceedings2019029179.

[4] Robison, M. (2020). Voice assistants have a gender bias problem. What can we do about it? *Forbes*. https://www.forbes.com/sites/matthewrobison/2020/03/10/voice-assistants-have-a-gender-bias-problem-what-can-we-do-about-it/ .

[5] Wold, J. B. (2006). Difficulties in learning English as a second or foreign language. *For the Learning of Mathematics, 26*(2), 27-31.

[6] Israel, G. D. (1992). Determining sample size. *University of Florida IFAS Extension Document PEOD-6*. Institute of Food and Agricultural Sciences (IFAS), University of Florida. https://edis.ifas.ufl.edu/publication/PD006 .

[7] Holzwarth, M., Janiszewski, C., & Neumann, M. M. (2006). The influence of avatars on online consumer shopping behavior. *Journal of Marketing, 70*(4), 33-48.

[8] Singh, P. (2020). *Implementing voice-activated interfaces with Python*. O'Reilly Media. (Book ISBN: 978-1492074793).

[9] Atal, B. S. and Rabiner, L. R. (1976). A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 24*(3), 201-212. https://doi.org/10.1109/TASSP.1976.1162821.

[10] Bapat, M., Gune, H., and Bhattacharyya, P. (2010). A paradigm-based finite state morphological analyzer for Marathi. *Proceedings of the ACL 2010 System Demonstrations*, 65-70. Association for Computational Linguistics (ACL). https://aclanthology.org/P10-4012.

[11] Staff, C. (2022). Women in computer science: Getting involved in STEM. *Computer Science.org*. https://www.computerscience.org/resources/women-in-computer-science/ .

[12] Schultz, T. and Waibel, A. (2001). Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication, 35*(1-2), 31-51. https://doi.org/10.1016/S0167-6393(00)00076-3.

[13] Carter, L. (2006). Why students with an apparent aptitude for computer science don't choose to major in computer science. *ACM SIGCSE Bulletin, 38*(1), 27-31. https://doi.org/10.1145/1124706.1121363.

[14] Eddy, S. R. (1996). Hidden Markov models. *Current Opinion in Structural Biology, 6*(3), 361-365. https://doi.org/10.1016/S0959-440X(96)80056-X.

[15] Faiz, Z., Srivastava, V. and Khoje, S. (2022). Virtual voice assistant for smart devices. *ECS Transactions, 107*(1), 4315-4323. Electrochemical Society. https://doi.org/10.1149/10701.4315ecst.

[16] Reddy, D. R. (1976). Speech recognition by machine: A review. *Proceedings of the IEEE, 64*(4), 501-531. https://doi.org/10.1109/PROC.1976.10175.

[17] Meyer, J., Dentel, L. and Meunier, F. (2013). Speech recognition in natural background noise. *PLoS ONE, 8*(11), e79279. https://doi.org/10.1371/journal.pone.0079279.

[18] Laroche, J. (2002). Time and pitch scale modification of audio signals. Book Chapter In M. Kahrs & K. Brandenburg (Eds.), *Applications of digital signal processing to audio and acoustics,* 169-203. Springer. https://doi.org/10.1007/978-1-4615-1863-0_7.

[19] Kim, D.-S., Lee, S.-Y. and Kil, R. M. (1999). Auditory processing of speech signals for robust speech recognition in real-world noisy environments. *IEEE Transactions on Speech and Audio Processing, 7*(1), 55-69. https://doi.org/10.1109/89.748214.

[20] Lima, L., Furtado, V., Furtado, E. and Almeida, V. (2019). Empirical analysis of bias in voice-based personal assistants. *Companion Proceedings of the World Wide Web Conference (WWW '19)*, 1085-1093. ACM. https://doi.org/10.1145/3308560.3317594 .

[21] Watson, S. (2019). The unheard female voice: Women are more likely to be talked over and unheeded. But SLPs can help them speak up and be heard. *The ASHA Leader, 24*(11), 44-51. American Speech Language Hearing Association (ASHA). https://doi.org/10.1044/leader.FTR3.24112019.44.

[22] Nasirian, F., Ahmadian, M. and Lee, O.-K. D. (2017). AI-based voice assistant systems: Evaluating from the interaction and trust perspectives. *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS)*, 1403-1412. University of Hawaii. https://doi.org/10.24251/HICSS.2017.170 .

[23] Bringsjord, S. and Schimanski, B. (2003). What is artificial intelligence? Psychometric AI as an answer. *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI '03)*, 887-893. Morgan Kaufmann Publishers. https://www.ijcai.org/proceedings/2003-1 .

[24] Yamazaki, K., Ueda, R., Nozawa, S., Kojima, M., Okada, K., Matsumoto, K., Ishikawa, M., Shimoyama, I. and Inaba, M. (2012). Home-assistant robot for an aging society.

Proceedings of the IEEE, 100(8), 2429-2441. https://doi.org/10.1109/JPROC.2012.2192086 .

[25] Zhang, N., Mi, X., Feng, X., Wang, X., Tian, Y. and Qian, F. (2019). Dangerous skills: Understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems. *2019 IEEE Symposium on Security and Privacy (SP)*, 1381-1396. IEEE. https://doi.org/10.1109/SP.2019.00081 .

[26] Diao, W., Liu, X., Zhou, Z. and Zhang, K. (2014). Your voice assistant is mine: How to abuse speakers to steal information and control your phone. *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones and Mobile Devices (SPSM '14)*, 63-74. ACM. https://doi.org/10.1145/2666624.2666628 .

[27] Zilberman, A. and Ice, L. (2021). Why computer occupations are behind strong STEM employment growth in the 2019–29 decade. *Computer, 54*(5), 11-15. IEEE. https://doi.org/10.1109/MC.2021.3057373 .