DESKTOP-BASED VIRTUAL ASSISTANT USING PYTHONBASED ON NATURAL LANGUAGE PROCESSING

Neha Shukla[a], Srijan Shahi[b], Sushant Kumar Pandey[c], Suraj Gupta[d]

Department of Computer Science, KIET Group of Institutions,

Delhi-NCR, Ghaziabad, India

neha.shukla@kiet.edu[a]

srijan.2024cs1083@kiet.edu[b]

sushant.2024cs1095@kiet.edu[c]

suraj.2024cs1034@kiet.edu[d]

Abstract - ***The incorporation of desktop virtual assistants into our everyday routines marks a noteworthy advancement in the field of human-computer interaction. This study examines the state of desktop virtual assistants, including their historical evolution, present situation, and possible future developments. From early AI and natural language processing projects like IBM's Shoebox to more complex systems like Siri and Google Assistant, desktop virtual assistants have come a long way. User's interactions with computers are revolutionized by their vast variety of features, which include task automation, natural language processing, voice recognition, and tailored suggestions. This article attempts to investigate the effects of the Assistant on user experience, productivity, and the wider computing future through in-depth investigation.***

Keywords **- Linguistic modeling, acoustic modeling, Artificial intelligence, and automation.**

I. Introduction

Almost all tasks are now digitalized in today's world. Voice searches have surpassed text searches. Web searches conducted via mobile devices have only recently surpassed those conducted via computer, and analysts predict that 50% of searches will be conducted via voice by 2024. Virtual assistants are turning out to be smarter than ever. Allow your intelligent assistant to handle your email. Detect intent, extract critical information, automate processes, and provide personalized responses. In recent years, several researchers have become interested in the recognition of human activities. The desktop's virtual assistant in Python is a software program that assists you with day-to-day tasks such as showing the weather report, creating reminders, making shopping lists, and so on.

They can respond to commands via text (as in online chatbots) or by voice. This system is intended for use on desktop computers. Virtual assistant software boosts user productivity by managing routine tasks and providing information from online sources. In this project, we propose a voice recognition system that recognizes human activities by utilizing an NLP algorithm. Voice is a form of communication in which users can communicate with one another. Automatic Speech Recognition (ASR), also known as voice recognition, recognizes spoken words and phrases and converts them to computer-readable formats. It accepts user input in the form of voice or text, processes it, and provides feedback to the user in a variety of ways, such as the action to be taken or the search result. As a result, distinguishing spoken words from background noise in audio is an additional challenge.

II. Literature Survey

[1] A thorough Systematic Literature Review (SLR) was conducted for this study to discover 21 distinct measures that are used to measure VA's user experience (UX). This study also presents the assessment standards for judging the operationalization rigor used in the creation of these scales. According to the study's analysis, the scales used to evaluate the user experience (UX) of virtual assistants (VAs) go beyond the conventional VUDA (value, usability, desirability, adaptability) principles and include cutting-edge ideas like anthropomorphism and machine personality. Despite some widespread and acknowledged standards, future VA UX researchers should also consider the differences in the stringent measures used during scale development. As a result, an overview is given together with recommendations for further research projects in the VA UX sector.

[2] This research on voice assistant attacks demonstrates that humans cannot understand the concealed vocal commands that control the VAs. Even though they are "hidden," hidden voice commands are audible. In this study, we create an entirely inaudible attack called DolphinAttack, which achieves inaudibility by modulating speech commands on ultrasonic carriers. The voice assistants can demodulate, recover, and—most importantly— interpret the modulated low-frequency audio commands by making use of the nonlinearity of the microphone circuits. We use well-known voice assistants like Siri, Google Now, S Voice, HiVoice, Cortana, Alexa, and others to confirm DolphinAttack. We demonstrate a few proof-of-concept assaults by inserting a series of inaudible voice instructions. These include using Siri to start a FaceTime call on an iPhone, turning on airplane mode on Google Now, and even tampering with the navigation system in an Audi car. We suggest defense-related hardware and software. By utilizing supported vector machines (SVM) to classify the audio, we confirm that DolphinAttack can be detected. We also propose re-designing voice assistants to make them resistant to attacks employing inaudible voice commands.

[3] With the vast network of IVA-enabled gadgets and cloud-hosted services from IVA and other developers, Figure 2 shows four potential attack vectors that could jeopardize user privacy and system security. Wiretapping an ecosystem for IVA. Sniffing the traffic between companion applications and the IVA can reveal the ecosystem's communication mechanisms even if the apps use encrypted network connections. For instance, we examined HTTPS requests and responses using packet interception tools to identify the APIs that are utilized to send and receive data to and from the IVA. Our investigation showed that not all network traffic is sent over a secure protocol when IVA-enabled devices and cloud-hosted services are communicating. An attacker posing as a user and using harmful voice commands to, for example, unlock a smart door to gain unlawful entry into a home or garage or place an order online without the user's knowledge, is a third security and privacy concern linked with IVAs. Even while some IVAs have a voice-training function to stop this kind of mimicry, the system may have trouble telling one voice from another. Consequently, an adversary with access to an IVA-enabled device may be able to deceive the system into believing that they are the true owner and engage in illicit or nefarious activities. Lastly, voices within the range of an IVA-enabled device may inadvertently be recorded and sent to the cloud, allowing third parties to listen in on private conversations. These parties may include businesses with authorized access to the stored data as well as hackers who may breach the database. Because of the possibility of unintentional recording, users may not have total control over their voice data.

[4] Two forms of voice assistant usage information search and task function—as well as five consumer values—social identity, convenience, personification, perceived utility, and perceived playfulness—are examined in this research. Using voice assistant technology, our findings contextualize and expand the TCV paradigm and provide empirical evidence for the relationship between consumption values. We discover that personification and social identity have a significant positive correlation with both playfulness and utility. Moreover, task function and information search are positively correlated with usefulness and playfulness. Furthermore,

there is a significant (and positive) moderating effect of trust and usage frequency on the relationship between voice assistant usefulness and utilization. These insights can be used by marketers and technology companies to create a range of voice-enabled services and applications that improve customer engagement and experience.

[5] Smart voice assistants (VAs) are being used extensively in smart home and industrial IoT systems, among other IoT systems, to offer voice-activated control services. However, an increasing number of attacks against virtual assistants (VAs) result in serious security issues because of the intricacy of the application environment and the variety of voice instructions. Due to voice development platforms' ability to access third-party voice abilities, adversaries might leverage ambiguous names in squatting attacks to get sensitive user data. Previous research examined how semantic misunderstanding in VA systems might be exploited for phonetic languages like English. However, the linguistic-model-guided fuzzing tool proposed by the previous study is not sufficient to perform semantic analysis on the VAs of Asian languages due to the semantic structural difference between phonetic English and symbol-based Asian languages, like Chinese. In this paper, we do a methodical investigation to assess the viability of voice misinterpretation attacks using semantic fuzzing against standard Asian language VAs.

We create Harmony-Fuzzer, a semantic fuzzing tool that uses abstracted fuzzing principles derived from speech error, dysfluency, and semantically related expression events in a Chinese corpus to guide the fuzzing process. To create statistical fuzzing models that allow the probability of fuzzing processing to govern the fuzzing space, we employ Bayesiannetworks.

[6] Voice-based personal assistants are becoming integral to our interaction with technology, evident by the rapid growth of smart speaker ownership in the U.S. With this reliance comes heightened concerns over security and privacy, prompting a surge in related research. This paper surveys the latest advancements in safeguarding personal voice assistants (PVAs), focusing on challenges linked to the acoustic channel. It delves into both potential threats and defensive strategies, covering well-established fields like voice authentication and emerging topics in acoustic security.

[7] The study anticipates an upswing in the adoption of digital voice assistants and devices featuring voice control. It categorizes systems with integrated voice user interfaces (VUIs) as voice assistants (VAs), defining VUIs as interfaces that facilitate interaction with technology through spoken language, eliminating the need for traditional input/output devices like keyboards, mice, and screens. VAs have permeated the consumer market, notably within smart devices, tablets, and PCs. However, there’s notable apprehension about their usage, particularly in Germany. The research examined the utilization of general-purpose VAs across specific technology-oriented demographics in Germany and Spain. Findings indicate that tech-savvy individuals are prolific VA users despite cultural variances. Both countries share concerns over privacy, command execution accuracy, and the need for VA enhancements. Many users in these regions worry about monitoring and data misuse. Media selection and voice transmission emerged as primary functions among these users. The study suggests that understanding the context of VA usage is crucial for enhancing user experience in human-computer interaction (HCI), though further data is necessary to fully grasp this context.

[8] The paper investigates security vulnerabilities in voice communication between users and virtual personal assistants (VPAs) like Amazon Alexa and Google Assistant. The study identifies two novel attacks: ‘voice squatting,’ where attackers use similarly pronounced or paraphrased names to hijack voice commands intended for legitimate skills, and ‘voice masquerading,’ where a malicious skill impersonates the VPA service or a legitimate skill to steal personal information. Through user studies and real-world tests on devices such as Amazon Echo andGoogle Home, these attacks were proven to be significant threats. The research contributed to the field by developing a squatting detector and a technique to automatically detect

masquerading attacks, which have been acknowledged by industry giants like Amazon and Google.

[9] The paper explores the implications of voice-based personal digital assistants (PDAs) on society, particularly focusing on ethical and legal challenges. While PDAs offer convenience and efficiency, they also raise concerns about privacy, data protection, and the potential for widening the digital divide. The discussion includes potential risks such as privacy breaches involving sensitive personal data, user engagement issues, and the impact on e-government services. The article also considers the role of big tech companies in this ecosystem and the need for robust feedback design to improve user experience. It highlights the importance of adhering to regulations like the General Data Protection Regulation (GDPR) to safeguard users' rights in the evolving landscape of internet services and computing devices.

[10] In modern society, various types of attacks are threatening security. In the field of voice, detecting voice spoofing attacks is an important issue. Several strategies for detecting speech replay or synthesis assaults employing loudspeakers have been developed. Liu et al. offered wearable technologies to detect voice liveness, such as spectacles, headphones, or necklaces. They reached about 97% accuracy in identifying liveness when utilizing headphones. To test voice liveness, Zhang et al. tracked unique articulatory motions using sound wave reflection techniques. They reached a 99.9% accuracy by having users physically hold their gadgets near their ears. Blue et al. employ sub-bass overexcitation and low-frequency signal characteristics to detect electronic speakers, obtaining 100% TAR and 1.72 FRR in calm environments. These techniques, like any other speech biometric-based technology, will all suffer considerable accuracy losses when subjected to background noise variations and environmental changes. Furthermore, merging numerous sophisticated models and characteristics necessitates extensive computational resources that are unsuitable for practical usage.

[11] The voice user interface (VUI) feedback mechanism is a key factor affecting the voice interaction's waiting experience. The feedback time of most available voice interfaces is fixed or decided by the processing time of hardware and software, which has not been designed and cannot offer users a good interaction experience. In this paper, the speech rate of user-machine voice interaction is collected through prototype experimentation. Besides, users' time perception of different voice interfaces' feedback time settings is studied based on time psychology theories. Moreover, users' emotional changes are described after a specific feedback time with the distribution of two-dimensional arousal-valence emotion space. Users' time perception and subjective emotions are differently influenced by different VUI feedback times. The experimental results show that 750 ms is the optimal VUI feedback time point at which the best users' subjective feelings and psychological experiences are reached, and the threshold limit time spent by users in waiting for the VUI feedback is 1,850 ms which will lead to user emotions with low levels of arousal and valence after being exceeded. Based on that, a linear regression model is proposed to define the optimal feedback time of VUI. The user experience VUI research results show that the calculated feedback time parameters can make users produce time perception in line with their expectations in interacting with voice interfaces. Verification of Correlation Between Time Perception and Speech Rate The usability of the linear regression model was experimentally verified in this study. First, the user groups of intelligent voice assistants in China were described according to the 2019 statistics entitled "Research Report on Enterprise Cases of China Intelligent Voice Assistant." A total of 20 teachers and students (10 males and 10 females, aged from 20 to 35) were selected from a university for this experiment. The subjects have the following characteristics: normal hearing, common and clear pronunciation of Mandarin, and no speech abnormality.

Table 1: The summary table of the reference mentioned in the Literature Review

S. No.	Author (Publishing Year)	Methodology	Remarks
1	Lawal Ibrahim Dutsinma Faruk, M. D. Babakerkhell, P. Mongkolnam [1]	To study of measure the user experience of voice assistants, a literature review gave 21 individual scales.	The 21 scales explain the different dimensions of user experience in an assistant scenario with a difference between Graphic interface and assistants.
2	C. Yan, G. Zhang, X. Ji, T. Zhang, T. Zhang, and W. Xu [2]	An attack that modulates input voice on an ultrasonic carrier to achieve inaudibility which can't be handled by a human being called a Dolphin Attack.	More no. of Inaudible voice attacks include activating Siri to call on iPhone and other existing assistance can be overcome by enhancing the microphone and Inaudible voice command cancellation.
3	S. Malodia, N. Islam, P. Kaur,and A. Dhir,[3]	Based on a different framework or mixed method approach that consists of interviewing experts, consumers, and active users of the assistant based on 5 consumption values and two types of virtual assistants.	Social identity and personification have a strong positive association with both usefulness and playfulness. In addition to this, usefulness and playfulness are positively associated with information search and task function.
4	H. Chung, M. Iorga, J. Voas, and S. Lee [4]	If an application uses a corrupted network connection Sniffing traffic between the application and an Intelligent voice assistant can expose the ecosystem communication mechanism	Malicious Voice commands, unsafe environment, unintentional voice recordings either accidentally or intentionally and transmitted to the cloud, give chance to hackers to do suspicious tasks.
5	J. Mao, Z. Liu, Q. Lin, and Z.Liang [5]	Voice development platforms allow third-party voice skills tobe accessed, adversaries can obtain users' private information by using confusing names.	with the usage of Bayesian networks to formulate a fuzzing model statistically and designing malicious skills to verify the feasibility of empirical attacks.
6	K. N. Lam, L. H. Nguy, V. L. Leand J. Kalita [31]	A Transformer-based model for diacritic restoration	It is two interconnected chatbots, both utilizing Transformers: a) a closed-domain chatbot trained on a comprehensive b) an open-domain chatbot trained on a vast movie dialog dataset.

7	Wenbin Huang; Wenjuan Tang; Hongbo Jiang; Jun Luo; Yaoxue Zhang [23]	a quasi-Gaussian distribution (QGD)-based defense scheme	Voice communication and automatic speech verification (ASV) through smart devices face vulnerabilities from voice impersonation (VI) attacks, where attackers mimic a target's voice to deceive human hearing or fool ASV systems.

III. Proposed methodology

1. User Interaction: The user and virtual assistant may now communicate with one other at this point. When a user gives an aural input, such as spoken instructions or inquiries, the procedure is started. A computer-connected microphone or a microphone integrated into a gadget like a smartphone or smart speaker are two examples of sources of audio input. In addition to priming the system to process user input, this stage establishes the context for the interaction.

2. Input Processing: Transforming the aural information into a format that the system can comprehend, and process further is the aim of input processing. Text transcription of spoken words is accomplished using speech recognition technologies. To detect individual words and translate them into written text, the audio stream must be analyzed. Sophisticated algorithms, which are frequently based on deep learning approaches, are used by modern voice recognition systems to obtain high transcription accuracy from spoken language. The transcription of audio input into text facilitates system manipulation and analysis of user input.

3. Text Processing: A crucial part of input processing is speech recognition, which specializes in translating spoken words into text. In this stage, language modeling is used to interpret the word order, and acoustic modeling is used to comprehend voice patterns. Speech recognition systems frequently employ methods like Deep Neural Networks (DNNs) and Hidden Markov Models (HMMs) to accomplish accurate transcription. To make sure that the system accurately interprets user input and can respond appropriately, voice recognition accuracy is crucial.

4. NLP Module: Now that the user has provided input in text format, text processing methods are used to examine and comprehend its content. Tokenization (splitting the text into individual words or tokens), part-of-speech tagging (determining the grammatical category of each word), and syntactic parsing (examining the grammatical structure of sentences) are a few examples of text processing activities.
   Text processing may also entail methods like sentiment analysis, which assesses the text's emotional tone, and named entity identification, which recognizes the names of individuals, locations, businesses, etc. Extraction of pertinent characteristics and information from user input is the aim of text processing; these features and information will be utilized in later stages to interpret user intent and produce relevant answers.

5. NLP Module: The goal of the artificial intelligence field of natural language processing (NLP) is to empower machines to comprehend, interpret, and produce human language. In this stage, NLP methods are used to further analyze and understand the text that was processed from the user's input.
   Beyond basic word-for-word comprehension, natural language processing (NLP) enables the system to comprehend the semantics (meaning) and pragmatics (context) of user input. NLP techniques include discourse

analysis (understanding the flow of a conversation), semantic analysis (interpreting meaning), and syntactic analysis (parsing sentence structure). NLP is essential to the virtual assistant's ability to understand human intent, deal with context and ambiguity, and provide relevant answers.

6. Determination of inquiry/Action: In this stage, the system determines whether the user's input is an instruction to execute an action or a request for information (inquiry) depending on how it interprets the input.
   In the event that the input is identified as a question or inquiry, the system will create a response by doing certain algorithmic operations or contacting pertinent information sources.
   The system carries out the desired job, which may entail communicating with other software programs or carrying out actions on the user's behalf if the input is recognized as a command or action. The system uses Text-to-voice (TTS) technology to synthesize the output into voice format after producing a response or performing the desired action, guaranteeing that the user hears the response.
7. Using the Google Text-to-voice (gTTS) API or a comparable TTS engine, the system transforms the produced text answer into synthesized voice in the last stage.
   With the use of algorithms, TTS technology creates speech that sounds natural while reading printed text. The technology produces the right intonation, rhythm, and pronunciation.
   The user is then presented with an aural output that translates the system's reaction or feedback when the synthesized speech is played back to them via speakers or headphones.
   TTS is necessary to provide smooth spoken language communication and to make the virtual assistant's interactions with the user more intuitive and natural.

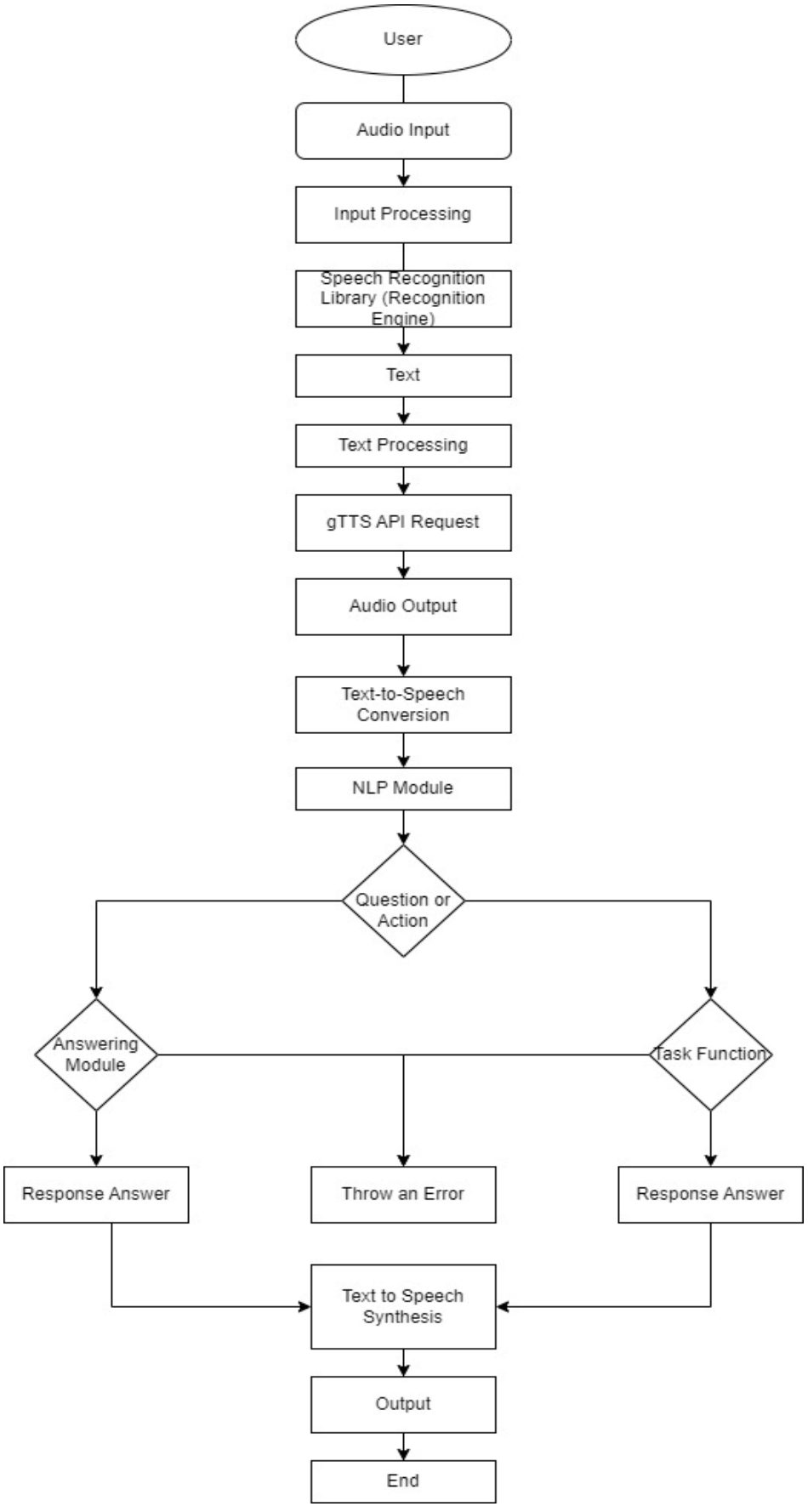

Figure 1. Flow chart for Desktop Virtual Assistant

IV. Conclusion

Virtual Assistants for the desktop that use Python are a very effective way to organize your schedule. Today, numerous Smart Personal Digital Assistant applications are available for a variety of device platforms. Because they have access to all of your Smartphone's resources, these new Software Applications outperform PDA devices. Because they are more portable and can be used at any time, virtual assistants are more dependable than human personal assistants. Because they have access to the internet, they have access to more information than any other assistant. The Python-based virtual assistant on the desktop is dependable and provides information in a user-friendly manner. and avoids the irrelevant info like ads. Syscall, in which a computer program requests a service from the kernel of the operating system on which it is executed. API call, system call, content extraction is interconnected to the python backend and from python backend, the information is passed to the text to speech module which converts the text data into speech. And the speech is returned to the user on his requirements using speakers.

V. References

[1.] L. I. D. Faruk, M. D. Babakerkhell, P. Mongkolnam, V. Chongsuphajaisiddhi, S. Funilkul, and D. Pal, "A Review of Subjective Scales Measuring the User Experience of Voice Assistants," in IEEE Access, vol. 12, pp. 14893-14917, 2024, doi: 10.1109/ACCESS.2024.3358423.

[2.] C. Yan, G. Zhang, X. Ji, T. Zhang, T. Zhang and W. Xu, "The Feasibility of Injecting Inaudible Voice Commands to Voice Assistants," in IEEE Transactions on Dependable and Secure Computing, vol. 18, no. 3, pp. 1108-1124, 1 May-June 2021, doi: 10.1109/TDSC.2019.2906165.

[1.] H. Chung, M. Iorga, J. Voas and S. Lee, "“Alexa, Can I Trust You?”," in Computer, vol. 50, no. 9, pp. 100-104, 2017, doi: 10.1109/MC.2017.3571053.

[2.] S. Malodia, N. Islam, P. Kaur and A. Dhir, "Why Do People Use Artificial Intelligence (AI)-Enabled Voice Assistants?," in IEEE Transactions on Engineering Management, vol. 71, pp. 491-505, 2024, doi: 10.1109/TEM.2021.3117884.

[3.] J. Mao, Z. Liu, Q. Lin and Z. Liang, "Semantic-Fuzzing-Based Empirical Analysis of Voice Assistant Systems of Asian Symbol Languages," in IEEE Internet of Things Journal, vol. 9, no. 12, pp. 9151-9166, 15 June15, 2022, doi: 10.1109/JIOT.2021.3113645.

[4.] P. Cheng and U. Roedig, "Personal Voice Assistant Security and Privacy—A Survey," in Proceedings of the IEEE, vol. 110, no. 4, pp. 476-507, April 2022, doi: 10.1109/JPROC.2022.3153167.

[5.] A. M. Klein, M. Rauschenberger, J. Thomaschewski and M. J. Escalona, "Comparing Voice Assistant Risks and Potential with Technology-Based Users: A Study from Germany and Spain," in Journal of Web Engineering, vol. 20, no. 7, pp. 1991-2016, October 2021, doi: 10.13052/jwe1540-9589.2071.

[6.] N. Zhang, X. Mi, X. Feng, X. Wang, Y. Tian and F. Qian, "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), San Francisco, CA, USA, 2019, pp. 1381-1396, doi: 10.1109/SP.2019.00016.

[7.] V. Almeida, E. S. Furtado and V. Furtado, "Personal Digital Assistants: The Need for Governance," in IEEE Internet Computing, vol. 24, no. 6, pp. 59-64, 1 Nov.-Dec. 2020, doi: 10.1109/MIC.2020.3009897.

[8.] I. -Y. Kwak et al., "Voice Spoofing Detection Through Residual Network, Max Feature Map, and Depthwise Separable Convolution," in IEEE Access, vol. 11, pp. 49140-49152, 2023, doi:

10.1109/ACCESS.2023.3275790.

[9.] J. Wang, Y. Li, S. Yang, S. Dong and J. Li, "Waiting Experience: Optimization of Feedback Mechanism of Voice User Interfaces Based on Time Perception," in IEEE Access, vol. 11, pp. 21241-21251, 2023, doi: 0.1109/ACCESS.2023.3250278.

[10.] Y. Uğurlu, M. Karabulut and İ. Mayda, "A Smart Virtual Assistant Answering Questions About COVID-19," 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Istanbul, Turkey,2020, pp. 1-6,doi:10.1109/ISMSIT50672.2020.9254350.

[11.] P. Spachos, S. Gregori and M. J. Deen, "Voice Activated IoT Devices for Healthcare: Design Challenges and Emerging Applications," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 69, no. 7, pp. 3101-3107,July, 2022, doi :10.1109/TCSII.2022.3179680.

[12.] W. Huang, W. Tang, H. Jiang, J. Luo and Y. Zhang, "Stop Deceiving! An Effective Defense Scheme Against Voice Impersonation Attacks on Smart Devices," in IEEE Internet of Things Journal, vol. 9, no. 7, pp. 5304-5314, 1 April1, 2022, doi: 10.1109/JIOT.2021.3110588.

[13.] K. N. Lam, L. H. Nguy, V. L. Le and J. Kalita, "A Transformer-Based Educational Virtual Assistant Using Diacriticized Latin Script," in IEEE Access, vol. 11, pp. 90094-90104, 2023,doi: 10.1109/ACCESS.2023.3307635.