VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JNANASANGAMA, BELAGAVI -590 018, KARNATAKA

TECHNICAL SEMINAR REPORT

ON

"VOICE RECOGNITION BASED IOT HOME AUTOMATION SYSTEM"

Submitted in the partial fulfillment of requirement for the award of Degree

B.E. in Computer Science & Engineering

SEMINAR ASSOCIATE

JEEVAN H K	**4BD20CS040**

SEMINAR GUIDES

Dr. Abdul Razak M S Ph.D.
Associate Professor,
Department of CS&E,
B.I.E.T., Davanagere.

Prof. Swetha H U M.Tech
Assistant Professor,
Department of AI&ML,
B.I.E.T., Davanagere.

Department of Computer Science and Engineering

Bapuji Institute of Engineering and Technology

Davanagere-577004

2023-24

Bapuji Institute of Engineering and Technology

Davangere -577004

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that **JEEVAN H K** bearing USN **4BD20CS040** of **Computer Science and Engineering** department have satisfactorily submitted the Technical Seminar(18CS84) Report entitled **"VOICE RECOGNITION BASED IOT HOME AUTOMATION SYSTEM"** in the partial fulfillment of the requirements for the award of Degree of Bachelor of Engineering (B.E.) in Computer Science& Engineering, under the VTU during the academic year 2023-24.

Dr. Abdul Razak M S Ph.D.
Associate Professor
Seminar Guide

Prof. Swetha H U M.Tech.,
Assistant Professor
Seminar Co-Guide

Prof. Rahima B M.Tech.,
Assistant Professor
Seminar Co-Ordinator

Prof. Madhu N Hiremath M.Tech.,
Assistant Professor
Seminar Co-Ordinator

Dr. Nirmala C R Ph.D.
Head of Department

Dr. H B Aravind Ph.D.
Principal

Date:
Place: Davanagere

Bapuji Educational Association (Regd.)
Bapuji Institute of Engineering and Technology, Davangere-577004

Department of Computer Science and Engineering

Vision and Mission of the Department

VISION

To be a center-of-excellence by imbibing state-of-the-art technology in the field of Computer Science and Engineering, thereby enabling students to excel professionally and be ethical.

MISSION

M1	Adapting best teaching and learning techniques that cultivates Questioning and Reasoning culture among the students.
M2	Creating collaborative learning environment that ignites the critical thinking in students and leading to the innovation.
M3	Establishing Industry Institute relationship to bridge the skill gap and make them industry ready and relevant.
M4	Mentoring students to be socially responsible by inculcating ethical and moral values.

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

The graduates will be able to

PEO1	To apply the skills acquired in the discipline computer science and Engineering for solving the societal and industrial problems with apt technology intervention.
PEO2	To continue their career in industry/academia or to pursue higher studies and Research
PEO3	To become successful entrepreneurs, innovators and job creators to design and develop software products and services to meet the societal, technical and business challenges
PEO4	To work in diversified environment by acquiring leadership qualities with strong Communication skills accompanied by professional and ethical values

PROGRAM SPECIFIC OUTCOMES (PSOs)

PSO1	Analyze and develop solutions for problems that are complex in nature but applying the knowledge acquired from the core subjects of this program.
PSO2	To develop secure, Scalable, Resilient and distributed applications for industry and societal requirements.
PSO3	To learn and apply the concepts and construct of emerging technologies like Artificial Intelligence, Machine learning, Deep learning, Big Data Analytics, IOT, Cloud Computing, etc for any real time problems.

ACKNOWLEDGEMENT

Salutations to our beloved and highly esteemed institute**, "BAPUJI INSTITUTE OF ENGINEERING AND TECHNOLOGY"** for having well qualified staff and lab furnished with necessary equipment's.

I express my sincere thanks to our guides **Dr. Abdul Razak M S** and **Prof. Swetha H U** for giving us constant encouragement, support and valuable guidance throughout the course of project without whose guidance this project would not have been achieved.

I express whole hearted gratitude to **Dr. Nirmala C R,** who is our respectable HOD of Computer Science and Engineering Department. I wish to acknowledge her help who made our task easy by providing with her valuable help and encouragement.

I express wholehearted gratitude to our Seminar Coordinator **Prof. Rahima B** and **Prof. Madhu N Hiremath**. I wish to acknowledge them who made our task easy by providing with her valuable help and encouragement.

I also express my wholehearted gratitude to our principal, **Dr. H B Aravind** for his moral support and encouragement.

I would like to extend my gratitude to all the staff of **Computer Science and Engineering Department** for their help and support. I have benefited a lot from the feedback, suggestions given by them.

I would like to extend our gratitude to all our family members and friends for their advice and moral support.

JEEVAN H K 4BD20CS040

ABSTRACT

Systems with voice control are an attractive option for increasing technological integration, not only for people with little knowledge on technology or constrained Internet access, but also for people with certain disabilities. In addition, devices based on Alexa or Google Home provide an interesting alternative for interacting with Internet of Things (IoT) devices, but they usually rely on an Internet connection to a cloud server for their full operation. To address the previously mentioned issues, this article presents a solution based on Edge Computing and voice commands that carries out offline voice processing and that is able to interact with IoT-based systems. The proposed system performs local speech inference, providing a communication interface with IoT devices in a Bluetooth mesh, all in a fast way and without the need for an Internet connection. In addition, the proposed solution can be adapted easily for voice recognition of languages with few resources. Such a feature is demonstrated with the Galician language, which is spoken by less than 3 million people worldwide. In particular, different Automatic Speech Recognition (ASR) models based on three of the most popular ASR development frameworks (wav2vec2, Distil Hubert, Whisper) were developed to transcribe short speech and to translate it into IoT commands that perform specific home automation actions. Such models were fine-tuned for Galician with a corpus of approximately 20 hours and were evaluated in static and mobile opportunistic scenarios in terms of accuracy, energy consumption and latency on an embedded platform (that acts as an edge device) and on a cloud server.

CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This chapter will give a brief introduction on the topic and then will be listing of various applications, advantages of this topic and some disadvantages.

1.1 BRIEF ON TECHNOLOGY

The emergence of the Internet of Things (IoT) has digitized various aspects of daily life, offering comfort and control. However, this technological progress is often inaccessible to certain groups, like the elderly and those in developing countries, due to barriers like complex user interfaces. Voice recognition systems offer a solution, providing intuitive interaction with IoT devices.

Challenges persist in implementing voice recognition, such as reliance on constant internet connection and privacy risks. However, advancements like Automatic Speech Recognition (ASR) systems have made strides, with options like wav2vec2 offering near-human precision with less transcription data.

To address language diversity and grammar limitations, frameworks like wav2vec2 allow for expanded voice recognition technology. The article introduces a voice recognition system for Galician, tailored for the elderly population in rural areas, where access to technology is limited.

Key contributions of the article include describing a home IoT system utilizing Edge Computing and optimized speech recognition for low-resource languages, analyzing ML models for ASR on edge devices, proposing optimization techniques, and evaluating the system's performance. The article's structure involves reviewing voice assistant technologies, analyzing Natural Language Processing (NLP) improvements, exploring ML in Edge Computing, detailing system design, presenting experimental results, and concluding with key findings and conclusions.

1. Voice Recognition Systems:

Voice Recognition Systems encompass various technologies aimed at converting spoken language into text for voice-controlled interactions with devices:

Automatic Speech Recognition (ASR): ASR systems like Amazon's Alexa and Google Home employ sophisticated algorithms to analyze and interpret spoken language, converting it into text format. These systems enable users to interact with devices through voice commands, facilitating tasks such as setting reminders, playing music, or controlling smart home devices.

Local Processing: Edge computing plays a crucial role in enhancing voice recognition systems by enabling local processing of data near the point of generation, such as on IoT devices or smart speakers. This approach reduces latency by eliminating the need to send data to distant cloud servers for analysis. Moreover, local processing enhances privacy by minimizing the exposure of sensitive data to external networks.

Wav2vec2: Developed by Facebook AI Research, Wav2vec2 represents a significant advancement in ASR technology. This framework utilizes self-supervised learning techniques, where the model learns from the input data without the need for extensive labeled datasets. Wav2vec2 can achieve high accuracy in speech recognition while requiring fewer labeled data for training, making it more accessible and efficient compared to traditional ASR approaches.

2. Language Availability:

Language Availability in voice recognition systems has been addressed through various approaches:

Wav2vec2: The Wav2vec2 framework has played a crucial role in expanding voice recognition capabilities to languages with limited labeled data. Traditional ASR systems typically require extensive labeled datasets for training, which may not be available for many languages. Wav2vec2 utilizes self-supervised learning techniques, enabling it to learn from unlabeled data and achieve high accuracy in transcription. This has made voice recognition technology more accessible to diverse linguistic communities by reducing the reliance on labeled data.

Keyword-Based Grammar: Instead of relying on exact transcriptions, some voice recognition systems utilize keyword-based grammar approaches. This method involves defining grammars around specific keywords or phrases that users are likely to use when interacting with devices. By focusing on the meaning of commands rather than the exact wording, keyword-based grammar enables more intuitive and natural language interactions with devices. Users can communicate with devices using variations of commands that feel more natural to them, enhancing the overall user experience.

3. Machine Learning (ML) Models:

Machine Learning (ML) Models play a crucial role in various voice recognition and natural language processing tasks:

Bidirectional Encoder Representations from Transformers (BERT): BERT is a highly influential pre-trained language model in the field of Natural Language Processing (NLP). It utilizes a transformer architecture to understand contextual meaning in text by considering the surrounding words. BERT models are capable of capturing intricate language patterns and nuances, making them suitable for a wide range of NLP tasks such as text classification, sentiment analysis, and question-answering.

DistilBert: DistilBert is a lighter and more compact version of the BERT model. Despite its reduced size, DistilBert retains much of the language understanding capabilities of its larger counterpart. By employing distillation techniques during training, DistilBert achieves a smaller model size and faster inference speed while preserving a significant portion of BERT's performance. This makes DistilBert particularly suitable for deployment on edge devices with limited computational resources, where efficiency and speed are essential considerations.

4. Edge Computing and Green IoT:

Edge Intelligence (EI): This approach involves processing data locally on edge devices, reducing reliance on centralized cloud servers and improving response times, critical for real-time applications like voice recognition.

Artificial Intelligence (AI) Acceleration: Specialized hardware like Application-Specific Integrated Circuits (ASICs), such as Google's Tensor Processing Unit (TPU), significantly enhance the performance of edge devices in executing ML models.

5. IoT Communication:

IoT Communication methods are crucial for facilitating data exchange and connectivity among devices:

Bluetooth Mesh: Bluetooth Mesh is a networking topology utilized in IoT applications, where devices form a mesh network to communicate with each other. Unlike traditional point-to-point connections, Bluetooth Mesh allows devices to relay data through multiple hops, enabling widespread coverage and scalability without the need for direct connections to a central server. This approach is particularly advantageous in IoT scenarios where devices are distributed across a large area or where infrastructure for centralized communication is limited. Bluetooth Mesh networks offer flexibility, resilience, and self-healing capabilities, making them well-suited for diverse IoT deployments.

6. Healthcare and Elderly Care Technologies:

Telemedicine: Leveraging technology to provide healthcare remotely, including video consultations, remote monitoring, and digital health platforms, particularly beneficial for elderly populations in remote areas.

IoT devices: Combined with voice-based assistants, IoT devices offer smart solutions for healthcare monitoring, medication reminders, and emergency assistance, enhancing the quality of life for elderly individuals.

7. Distributed Topologies:

In IoT environments, distributed topologies enable direct communication between devices, reducing dependence on centralized servers. This approach enhances network efficiency, scalability, and robustness, making it ideal for large-scale IoT deployments. Distributed topologies facilitate peer-to-peer communication, enabling devices to interact seamlessly without relying on a single point of failure. By distributing computational tasks across multiple nodes, distributed topologies optimize resource utilization and improve overall system performance.

8. Energy-Efficient Hardware:

Embedded Architectures: Embedded architectures are hardware systems optimized for specific tasks with low power consumption. These systems are commonly used in IoT devices deployed in remote or resource-constrained environments where energy efficiency is critical. Embedded architectures offer a balance between computational power and energy consumption, enabling IoT devices to operate efficiently on limited power sources such as batteries or solar panels. By minimizing energy usage, embedded architectures prolong device lifespan and reduce maintenance requirements, making them well-suited for long-term IoT deployments in diverse settings.

1.2 APPLICATIONS

The below seven are the applications of the presented solution based on Edge Computing and voice commands for interacting with IoT-based systems are manifold

1.Accessibility Enhancement:

- The system's voice-controlled interface caters to individuals with limited technological knowledge by providing a natural and intuitive interaction method.
- For individuals with disabilities, such as motor impairments or visual impairments, voice commands offer an alternative means of interacting with IoT devices, enhancing accessibility.

2.Offline Voice Processing:

- Offline voice processing eliminates the reliance on an Internet connection, making the system suitable for environments with limited or unreliable internet access.
- This capability ensures consistent functionality regardless of network availability, making the solution more robust and reliable.

3.Language Adaptability:

- The system's ability to adapt to languages with limited resources, exemplified by its support for the Galician language, enhances inclusivity by catering to minority language speakers.
- Adapting the solution for other languages with limited resources expands its reach and usability in diverse linguistic communities.

4.Home Automation:

- The system enables seamless integration with IoT devices for home automation, allowing users to control various aspects of their smart homes using voice commands.

- Users can perform tasks such as adjusting lighting, controlling appliances, setting thermostats, and managing security systems through voice interactions, enhancing convenience and efficiency.

5.Real-Time Response:

- With fast local speech inference and communication with IoT devices in a Bluetooth mesh network, the system ensures real-time response to voice commands.
- Minimal latency in executing actions enhances user experience and responsiveness, making the interaction with IoT devices seamless and efficient.

6.Energy Efficiency:

- The implementation of the solution on embedded devices optimizes energy consumption, making it suitable for deployment in resource-constrained environments.
- Energy-efficient operation prolongs device battery life and reduces overall energy consumption, ensuring sustainability and cost-effectiveness.

7.Scalability:

- The system's architecture, leveraging Edge Computing principles, enables scalable deployment across diverse IoT environments.
- Distributed communication and edge processing facilitate the integration of a large number of devices while maintaining performance and reliability, ensuring scalability as the IoT ecosystem expands.

1.3 ADVANTAGES:

The below advantages of the presented solution based on Edge Computing and voice commands for interacting with IoT-based systems are manifold.

Increased Technological Integration: Voice-controlled systems enhance technological integration, particularly for individuals with limited technological knowledge or internet access. This makes technology more accessible and user-friendly for a wider range of users.

Accessibility for People with Disabilities: Voice-controlled systems provide an alternative interaction method for individuals with disabilities, such as motor impairments or visual impairments, who may have difficulty using traditional interfaces.

Offline Voice Processing: The system performs offline voice processing, eliminating the need for a constant internet connection. This ensures consistent functionality even in environments with limited or unreliable internet access.

Language Adaptability: The solution can be easily adapted for voice recognition in languages with limited resources, catering to minority language speakers and promoting inclusivity in linguistic diversity.

Local Speech Inference: By performing local speech inference, the system reduces latency and dependence on cloud servers, leading to faster response times and improved user experience.

Bluetooth Mesh Communication: Communication with IoT devices via Bluetooth mesh enables seamless interaction without the need for an internet connection. This enhances connectivity and interoperability in IoT environments.

Adaptability to Edge Devices: The solution is optimized for edge devices, such as Raspberry Pi, making it suitable for deployment in resource-constrained environments or remote locations.

Energy Efficiency: Optimized speech processing and local inference contribute to energy-efficient operation, prolonging device battery life and reducing overall energy consumption.

1.3 LIMITATIONS:

The below limitations of the presented solution based on Edge Computing and voice commands for interacting with IoT-based systems are manifold.

Limited Language Support: While the solution can be adapted for languages with limited resources, it may still face challenges in supporting a wide range of languages, especially those with fewer speakers and linguistic resources.

Accuracy and Performance Trade-offs: Achieving high accuracy in speech recognition on edge devices may require trade-offs in terms of model complexity and performance. Balancing accuracy with resource constraints can be challenging.

Hardware Requirements: Certain features, such as real-time speech inference, may require hardware acceleration or more powerful edge devices, which could increase deployment costs or hardware requirements.

Scalability Challenges: Scaling the solution to accommodate a large number of IoT devices or users may pose challenges in terms of system performance, management, and maintenance.

Security and Privacy Concerns: Storing and processing voice data locally raises security and privacy concerns, especially regarding data encryption, user authentication, and protection against unauthorized access.

Dependency on Local Infrastructure: The system's reliance on local infrastructure, such as Bluetooth mesh networks, may limit its usability in certain environments or scenarios where such infrastructure is unavailable or unreliable.

Maintenance and Updates: Ensuring the reliability and security of the system may require regular maintenance and updates, which could be challenging in remote or resource-constrained environments.

CHAPTER 2

LITERATURE SURVEY

This chapter will present the different papers by different authors who have been researched on the similar topics.

1. United Nations. "World Population Ageing 2020 Highlights". Accessed: Feb. 2023. [Online]. Available: https://www.un.org/development/desa/pd/news/world-population-ageing-2020-highlights

The Department of Economic and Social Affairs of the United Nations Secretariat is a vital interface between global policies in the economic, social and environmental spheres and national action. The Department works in three main interlinked areas: (i) it compiles, generates and analyses a wide range of economic, social and environmental data and information on which States Members of the United Nations draw to review common problems and take stock of policy options; (ii) it facilitates the negotiations of Member States in many intergovernmental bodies on joint courses of action to address ongoing or emerging global challenges; and (iii) it advises interested Governments on the ways and means of translating policy frameworks developed in United Nations conferences and summits into programmes at the country level and, through technical assistance, helps build national capacities. The Population Division of the Department of Economic and Social Affairs provides the international community with timely and accessible population data and analysis of population trends and development outcomes for all countries and areas of the world. To this end, the Division undertakes regular studies of population size and characteristics and of all three components of population change (fertility, mortality and migration). Founded in 1946, the Population Division provides substantive support on population and development issues to the United Nations General Assembly, the Economic and Social Council and the Commission on Population and Development. The Division leads or participates in various interagency coordination mechanisms of the United Nations system. It also contributes to strengthening the capacity of Member States to monitor population trends and to address current and emerging population issues.

2. P. Kaur, P. Singh, and V. Garg, ‘‘Speech recognition system; challenges and techniques,’’ Int. J. Comput. Sci. Inf. Technol., vol. 3, no. 3, pp. 3989–3992, 2012.

Pattern matching is an emerging subject for doing identification of various objects including speech and its parts. Initially these pattern matching techniques has been limited to identifying the speech patterns of the words that are separated with enough silence. It is easy to segment each word and extract its features which machine algorithm like Neural Network can learn and simulate, but if a sufficient interval of silence is not there, then there are connected words between each pause based on persons rate of speech, if for example person has spoken 5 types of words the connected combination might reached to 120 which would ultimately lead to a great challenge for pattern matching algorithm. In this paper we are reviewing such methodologies.

3. S. F. N. Zaidi, V. K. Shukla, V. P. Mishra, and B. Singh, ‘‘Redefining home automation through voice recognition system,’’ in Emerging Technologiesin Data Mining and Information Security (Advances in Intelligent Systems and Computing), vol. 1300, A. E. Hassanien, S. Bhattacharyya, S. Chakrabati, A. Bhattacharya, and S. Dutta, Eds. Singapore: Springer, 2021, doi:10.1007/978-981-33-4367-2_16.

The world is evolving into a digitally advanced environment. Transformation is a constantly evolving process where Robotic Process Automation or RPA came into the process of renovation. RPA has recently become a valuable tool in banking and financial institutions. RPA has shown a lot of various benefits for different organizations. The primary intention of Robotic Process Automation in banking is to reduce repetitive tasks in the bank. In banking and various other organizations, RPA has helped reduce the operational costs by 30% - 70%; RPA helps reduce the workforce by employing Bot workers in charge, which later saves the operating costs and increases efficiency and accuracy of the tasks. Lenders are regularly facing pressure to reduce the prices as well as to reduce and save time. Lender hence switches into automation for better efficiency and accuracy of service. With automation bots, lenders can automate loan processing by collecting customer information, loan approval, loan monitoring, and automatic loan pricing. This can be achieved with the help of rule-based software bots. Also, many of the lenders do some part of the process partially automated and some part manually. Banks and financial institutions are switching to automation and training to stay on top of the latest security developments. This helps to keep an eye on the evolving trends in the payment space. Fraud for instance, is an ongoing threat. Banks, insurance companies, and other financial institutions employ this new age of RPA technology. This is to identify and counter frauds pulling data from multiple service lines instead of creating a bunch of economic macros. The paper talks about how RPA can mitigate fraud risks through various methods such as reassessing current processes, eliminating human errors, enhanced trade monitoring, automated threat detection, and searching for anomalies and much more.

4. “Voice Recognition Tech Privacy and Cybersecurity Concerns. Accessed”: Feb. 2023.

A new report released by Global Market Insights, Inc. last month estimates that the global market valuation for voice recognition technology will reach approximately $7 billion by 2026, in main part due to the surge of AI and machine learning across a wide array of devices including smartphones, healthcare apps, banking apps and connected cars, just to name a few. Whether performing a quick handsfree search on your phone or car command while driving, voice recognition technology has enhanced the effortlessness of consumer use. Particularly in the wake of the COVID-19 pandemic, companies that may never have considered voice-recognition technology are now rethinking their employee access control systems, and considering touchless authorization technologies, like voice recognition, as the main form of entry into their workspace, as opposed to fingerprint scanners or keypads that increase the risk of germs or virus spreading.

But while the ease and efficiency of voice recognition technology is clear, the privacy and security obligations associated with this technology cannot be overlooked. Voice recognition is generally classified as a biometric technology which allows the identification of a unique human characteristic (e.g. voice, speech, gait, fingerprints, iris or retina patterns), and as a result voice related data qualifies biometric information and in turn personal information under various privacy and security laws. For businesses that want to deploy voice recognition technology, whether for use by their employees to access systems or when manufacturing a smart device for consumers or patients, there are a number of privacy and security compliance obligations to consider.

5. J. Lau, B. Zimmerman, and F. Schaub, ‘‘Alexa, are you listening?’’ Proc. ACM Hum.-Comput. Interact., vol. 2, pp. 1–31, Nov. 2018.

Smart speakers with voice assistants, like Amazon Echo and Google Home, provide benefits and convenience but also raise privacy concerns due to their continuously listening microphones. We studied people's reasons for and against adopting smart speakers, their privacy perceptions and concerns, and their privacy-seeking behaviors around smart speakers. We conducted a diary study and interviews with seventeen smart speaker users and interviews with seventeen non-users. We found that many non-users did not see the utility of smart speakers or did not trust speaker companies. In contrast, users express few privacy concerns, but their rationalizations indicate an incomplete understanding of privacy risks, a complicated trust relationship with speaker companies, and a reliance on the socio-technical context in which smart speakers reside. Users trade privacy for convenience with different levels of deliberation and privacy resignation. Privacy tensions arise between primary, secondary, and incidental users of smart speakers. Finally, current smart speaker privacy controls are rarely used, as they are not well-aligned with users' needs. Our findings can inform future smart speaker designs; in particular we recommend better integrating privacy controls into smart speaker interaction.

6. A.-L. Georgescu, A. Pappalardo, H. Cucu, and M. Blott, ‘‘Performance vs. hardware requirements in state-of-the-art automatic speech recognition,’’ EURASIP J. Audio, Speech, Music Process., vol. 2021, no. 1, pp. 1–30, Jul. 2021.

The last decade brought significant advances in automatic speech recognition (ASR) thanks to the evolution of deep learning methods. ASR systems evolved from pipeline-based systems, that modeled hand-crafted speech features with probabilistic frameworks and generated phone posteriors, to end-to-end (E2E) systems, that translate the raw waveform directly into words using one deep neural network (DNN). The transcription accuracy greatly increased, leading to ASR technology being integrated into many commercial applications. However, few of the existing ASR technologies are suitable for integration in embedded applications, due to their hard constrains related to computing power and memory usage. This overview paper serves as a guided tour through the recent literature on speech recognition and compares the most popular ASR implementations. The comparison emphasizes the trade-off between ASR performance and hardware requirements, to further serve decision makers in choosing the system which fits best their embedded application. To the best of our knowledge, this is the first study to provide this kind of trade-off analysis for state-of-the-art ASR systems.

7. S. Gondi and V. Pratap, ‘‘Performance evaluation of offline speech recognition on edge devices,’’ Electronics, vol. 10, no. 21, p. 2697, Nov. 2021.

Deep learning–based speech recognition applications have made great strides in the past decade. Deep learning–based systems have evolved to achieve higher accuracy while using simpler end-to-end architectures, compared to their predecessor hybrid architectures. Most of these state-of-the-art systems run on backend servers with large amounts of memory and CPU/GPU resources. The major disadvantage of server-based speech recognition is the lack of privacy and security for user speech data. Additionally, because of network dependency, this server-based architecture cannot always be reliable, performant and available. Nevertheless, offline speech recognition on client devices overcomes these issues. However, resource constraints on smaller edge devices may pose challenges for achieving state-of-the-art speech recognition results. In this paper, we evaluate the performance and efficiency of transformer-based speech recognition systems on edge devices. We evaluate inference performance on two popular edge devices, Raspberry Pi and Nvidia Jetson Nano, running on CPU and GPU, respectively. We conclude that with PyTorch mobile optimization and quantization, the models can achieve real-time inference on the Raspberry Pi CPU with a small degradation to word error rate. On the Jetson Nano GPU, the inference latency is three to five times better, compared to Raspberry Pi. The word error rate on the edge is still higher, but it is not too far behind, compared to that on the server inference.

8. C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, ‘‘Applying wav2vec2.0 to speech recognition in various low-resource languages,’’ 2020, arXiv:2012.12121.

There are several domains that own corresponding widely used feature extractors, such as ResNet, BERT, and GPT-x. These models are usually pre-trained on large amounts of unlabeled data by self-supervision and can be effectively applied to downstream tasks. In the speech domain, wav2vec2.0 starts to show its powerful representation ability and feasibility of ultra-low resource speech recognition on the Librispeech corpus, which belongs to the audiobook domain. However, wav2vec2.0 has not been examined on real spoken scenarios and languages other than English. To verify its universality over languages, we apply pre-trained models to solve low-resource speech recognition tasks in various spoken languages. We achieve more than 20% relative improvements in six languages compared with previous work. Among these languages, English achieves a gain of 52.4%. Moreover, using coarse-grained modeling units, such as subword or character, achieves better results than fine-grained modeling units, such as phone or letter.

9. E. Guglielmi, G. Rosa, S. Scalabrino, G. Bavota, and R. Oliveto, ''Sorry, I don't understand: Improving voice user interface testing,'' in Proc. 37th IEEE/ACM Int. Conf. Automated Softw. Eng., Oct. 2022, pp. 1–12.

Voice-based virtual assistants are becoming increasingly popular. Such systems provide frameworks to developers on which they can build their own apps. End-users can interact with such apps through a Voice User Interface (VUI), which allows to use natural language commands to perform actions. Testing such apps is far from trivial: The same command can be expressed in different ways. To support developers in testing VUIs, Deep Learning (DL)-based tools have been integrated in the development environments (e.g., the Alexa Developer Console, or ADC) to generate paraphrases for the commands (seed utterances) specified by the developers. Such tools, however, generate few paraphrases that do not always cover corner cases. In this paper, we introduce VUIUPSET, a novel approach that aims at adapting chatbot-testing approaches to VUI-testing. Both systems, indeed, provide a similar natural-language-based interface to users. We conducted an empirical study to understand how VUI-UPSET compares to existing approaches in terms of (i) correctness of the generated paraphrases, and (ii) capability of revealing bugs. Multiple authors analyzed 5,872 generated paraphrases, with a total of 13,310 manual evaluations required for such a process. Our results show that, while the DLbased tool integrated in the ADC generates a higher percentage of meaningful paraphrases compared to VUI-UPSET, VUI-UPSET generates more bug-revealing paraphrases. This allows developers to test more thoroughly their apps at the cost of discarding a higher number of irrelevant paraphrases.

10. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, ‘‘Wav2vec 2.0: A framework for self-supervised learning of speech representations,’’ in Proc. Adv.Neural Inf. Process. Syst., vol. 33, 2020, pp. 12449–12460.

They show for the first time that learning powerful representations from speech audio alone followed by fine-tuning on transcribed speech can outperform the best semi-supervised methods while being conceptually simpler. wav2vec 2.0 masks the speech input in the latent space and solves a contrastive task defined over a quantization of the latent representations which are jointly learned. Experiments using all labeled data of Librispeech achieve 1.8/3.3 WER on the clean/other test sets. When lowering the amount of labeled data to one hour, wav2vec 2.0 outperforms the previous state of the art on the 100 hour subset while using 100 times less labeled data. Using just ten minutes of labeled data and pre-training on 53k hours of unlabeled data still achieves 4.8/8.2 WER. This demonstrates the feasibility of speech recognition with limited amounts of labeled data.1

CHAPTER 3

METHODOLOGY

The methodology chapter of the described article likely begins with outlining the research objectives and the specific aims of the proposed solution.

1. MAIN COMPONENTS OF THE PROPOSED SYSTEM

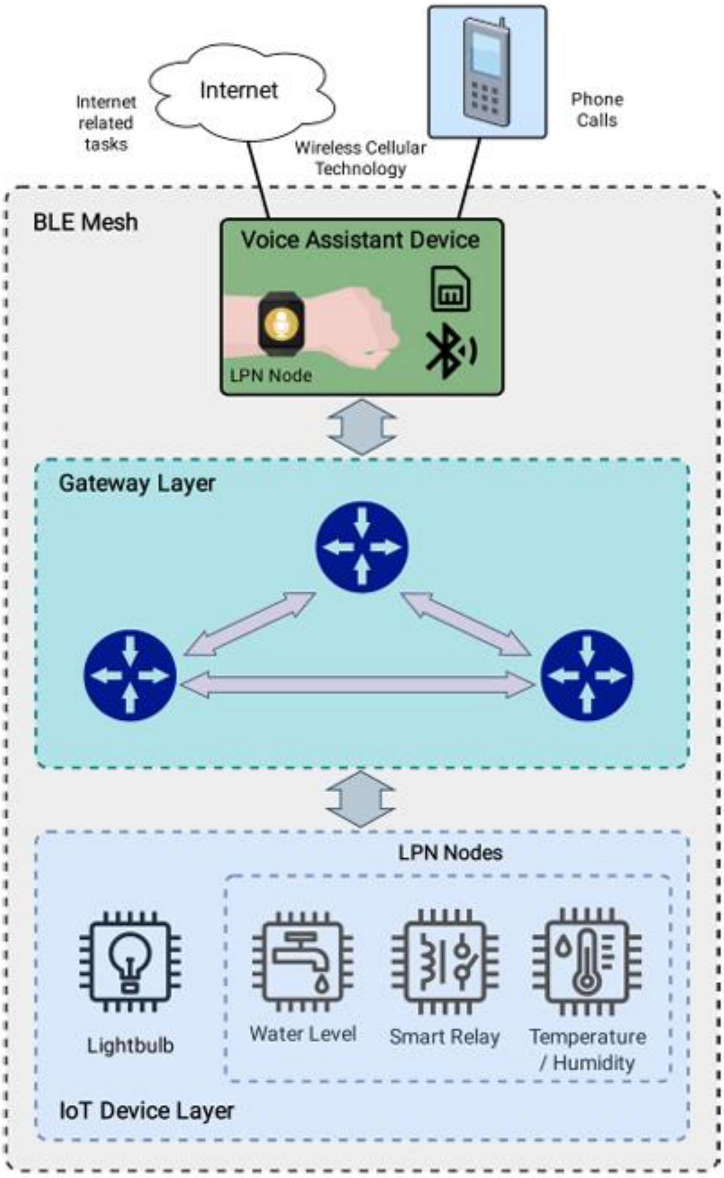


FIGURE 3.1. Methodology.

Above Figure 3.1 shows the general methodology of the proposed system. Specifically, the different layers of the developed system.

- **IOT DEVICE LAYER**

  The IoT device layer includes the different sensors/ actuators nodes of the system. It is worth noting that the layer includes nodes with constant energy sup ply that need fast response times while others do not have critical time-response requirements or have power supply restrictions (thus prioritizing energy saving over response time, becoming Low Power-consumption Nodes (LPN)).

- **GATEWAY LAYER**

  The gateway layer includes the different intermediate relay nodes, which provide the desired communications coverage to a house or building.

- **THE VAD LAYER**

  The VAD layer essentially includes the voice assistant node. Since such a node runs on batteries, it is considered as an LPN node. The VAD will interact with the nodes of the gateway and IoT device layers to perform voice induced actions. If necessary, it can also perform actions that involve using an Internet connection (e.g., check the weather, receive news, collect the user e-mail) or making calls through acellular network, but, nonetheless, it must be noted that the presented architecture has been con ceived and designed to be deployed in a house without any infrastructure with Internet access.

2. COMMUNICATIONS TECHNOLOGY SELECTION

The selection of communication technology for the proposed system involved analyzing various factors such as standardization, operating frequency, range, modulation scheme, encryption, topology, latency, battery lifetime, and cost. After thorough consideration, Bluetooth 5 was chosen for several reasons:

Improved Features: Bluetooth 5 offers better range, latency, bandwidth, and coexistence compared to its predecessor, making it suitable for the proposed system's requirements.

Backward Compatibility: Bluetooth 5 is backward compatible, allowing it to work alongside other Bluetooth versions, ensuring compatibility with existing devices.

Open and Widely Documented Standard: The Bluetooth standard is open and well-documented, facilitating more accurate and specific developments, which is advantageous for customization and integration.

Mesh Topology Support: Bluetooth 5 supports a mesh topology, which aligns well with the system's needs, as it requires a reduced network with simple messages without many hops.

Energy Efficiency: Bluetooth 5 is energetically efficient, which is crucial for battery-powered devices commonly used in home automation applications.

Adequate Response Times: Bluetooth 5 provides adequate response times, essential for applications such as lighting controls, where immediate response is crucial.

3.BLE MESH COMMUNICATIONS

Internal communications are performed with BLE Mesh. Such communications are based on topics: performing an action means publishing a certain value in a specific topic, and consulting the status of a device means subscribing to a specific topic. There are nodes connected to a continuous power supply that require a fast response (such as lights), where there are also LPN nodes that operate with batteries or limited power supplies that need to be in periodic sleep cycles (the information addressed to these nodes is stored in intermediate nodes). The BLE Mesh standard includes a feature called ‘‘friend node’’ that allows for storing information for the LPN nodes. However, friend nodes do not act as a distributed cache: when an LPN node loses its connection to the associated friend node, its cache is lost. This is not a problem for most LPNs, since they are static. Occasionally, in very specific cases, they may lose the connection to their friend node, in which case they will look for another one that is in range to store the cache. However, the VAD node, which is not an LPN node in the strict sense of the term ,as it is always listening for commands: from the point of view of the Bluetooth subsystem, it works as an LPN node, sleeping most of the time and waking up occasionally to consult the cache or when a voice command is provided. To overcome this limitation, in the developed system, when a LPN node needs to send a message to the VAD node, it sends it to all the friend nodes. Thus, when the VAD connects to any node, it will read the message and the rest of the duplicate messages will be discarded.

4. COMMUNICATIONS SUBSYSTEM OPTIMIZATION

The communications subsystem is responsible for facilitating communication with end devices for executing voice-commanded operations. The system's total latency is the sum of ASR model processing time and action execution time. Bluetooth Low Energy (BLE) Mesh, supported by Nordic boards, is chosen for its efficiency and wide support in IoT devices. Generic BLE Mesh models are used for communication, optimized for efficiency with reduced frame sizes. The behavior of the Voice Activity Detection (VAD) node differs from standard nodes, operating opportunistically to conserve power. Overall, the system prioritizes efficiency and simplicity in communication for effective voice command recognition and IoT device interaction.

$$t_{total} = t_{transcription} + t_{action} \tag{1}$$

$$t_{action} = t_{detection\_comm} + t_{sending} + t_{execution} \tag{2}$$

Equation 1 indicates the factors involved in the calculation of the total latency of the system. As it can be observed, it is calculated as the processing time of the ASR model to generate the transcript plus the time to perform the action. Equation 2 shows the different partial times involved in the calculation of the time to perform an action, which is com posed by the time that is needed to detect the particular command once the transcript is generated plus the time required to send it through the communications subsystem and the time consumed by the destination node to receive it and to execute the action.

5. SPEECH-RECOGNITION FRAMEWORK SELECTION.

The evolution of speech recognition frameworks has led to the emergence of self-supervised learning techniques, notably exemplified by wav2vec2, HuBERT, and Whisper.

These frameworks leverage self-supervised learning to generate meaningful representations of speech audio, followed by fine-tuning on transcribed speech, resulting in improved transcription accuracy, particularly for languages with limited resources.

Wav2vec2, while self-supervised, is pretrained on clean, read speech from audiobooks, potentially limiting its accuracy in processing noisy, conversational audio. In contrast, Whisper is trained in a supervised fashion on a vast corpus of multilingual speech data, albeit weakly supervised, thus offering greater scale and diversity in training data.

For the edge-intelligent speech recognition solution described in the article, wav2vec2, HuBERT, and Whisper architectures were tested, adapted, and evaluated to meet the requirements and constraints of performing machine learning tasks on edge devices.

6.SPEECH-RECOGNITION FRAMEWORK OPTIMIZATION AND FINETUNING

The system utilizes wav2vec2, a distilled version, and Whisper models for Voice Activity Detection (VAD) adapted to Galician language and mobile devices. Text distance algorithms detect keywords for command execution due to the lack of Galician grammar corpuses. Various optimization techniques like quantization, pruning, and model distillation are employed to reduce model size and improve inference speed. Inference runtime is optimized using mobile ML frameworks and acceleration mechanisms. Success rate, along with Word Error Rate (WER), measures transcription accuracy, with algorithms comparing transcribed text with predefined keywords for command detection.

Equation 3 defines the detection of keywords between strings a and b using the Levenshtein

$$det(a, b) = \frac{sim(a, b)}{lev(a, b)} \qquad (3)$$

distance and similarity.

$$sim(a, b) = max(|a|, |b|) - lev(a, b) \qquad (4)$$

Equation 4 shows the expression used for obtaining the similarity between two strings a and b.

$$lev(a,b) = \begin{cases} |a| & if\ |b| = 0, \\ |b| & if\ |a| = 0, \\ lev(tail(a), tail(b)) & if\ a[0] = b[0], \\ 1 + min \begin{cases} lev(tail(a), b) \\ lev(a, tail(b)) \\ lev(tail(a), tail(b) \end{cases} & otherwise, \end{cases} \tag{5}$$

Equation 5 shows the calculation of Levenshtein distance between two strings (a,b), where tail(x) is a substring of all but the first character of x. As it can be observed, the second block of Equation 5 represents the deletion, insertion and substitutions performed for all substrings of (a,b).

$$det(a,b) = 0 \vee det(a,b) \geq \alpha \implies detection \tag{6}$$

Equation 6 defines the condition necessary for a detection, where α is a predefined threshold.

$$det(a,b) = \frac{2 \times (|a| + |b|)}{M} \tag{7}$$

A simpler approach is applied for the similarity calculation with the LCS algorithm. Keyword detection is performed for such an algorithm through Equation 7 for two words a and b, being M the number of matches between the longest common subsequence of strings a and b.

$$det(a,b) \geq \alpha \implies detection \quad | \quad \alpha \in [0,1] \subset \mathbb{R} \tag{8}$$

Finally, Equation 8 shows the condition necessary for a detection when using the LCS algorithm, where α is a pre defined threshold normalized between 1 and 0.

CHAPTER 4

EXPERIMENTAL RESULTS

The results section presents the findings and outcomes obtained from the experimentation and analysis conducted.

4.1 RESULTS

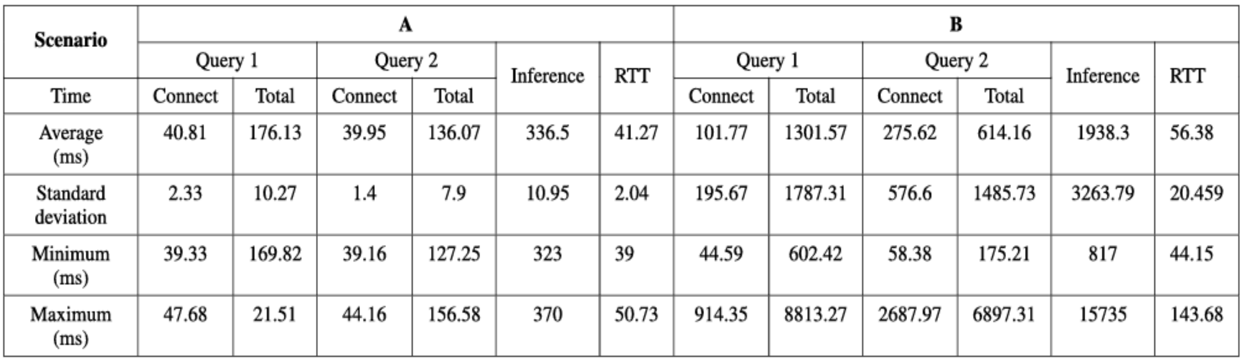
This is an evaluation of a speech inference system designed to interpret voice commands and perform corresponding actions. The system's primary objective is not limited to transcribing speech but extends to understanding and executing commands effectively. To assess its performance, we conducted tests using 25 voice commands, each lasting 5 seconds, representing various categories such as controlling lights, heating, notifications, alarms, volume adjustment, and making phone calls. These commands were expressed naturally and spontaneously, reflecting real-world speech patterns to evaluate the system's robustness.

The system relies on predefined keywords, totaling 28, to determine the intended command within the utterance. For instance, keywords like "speak" and "louder" in the phrase ‘‘Non escoito ben, fala máis alto’’ indicate the command falls under the "increase/decrease volume" category. The tested commands spanned eight different categories, offering a diverse set of tasks for the system to handle.

Performance evaluation focused on two key metrics: effectiveness and latency. Effectiveness measured the accuracy of correctly identifying the command, while latency quantified the time taken for the system to recognize and act upon the command.

The below results provide insights into the system's ability to understand and respond to voice commands in real-world scenarios, considering factors such as natural language variation, keyword detection accuracy, and task execution efficiency.

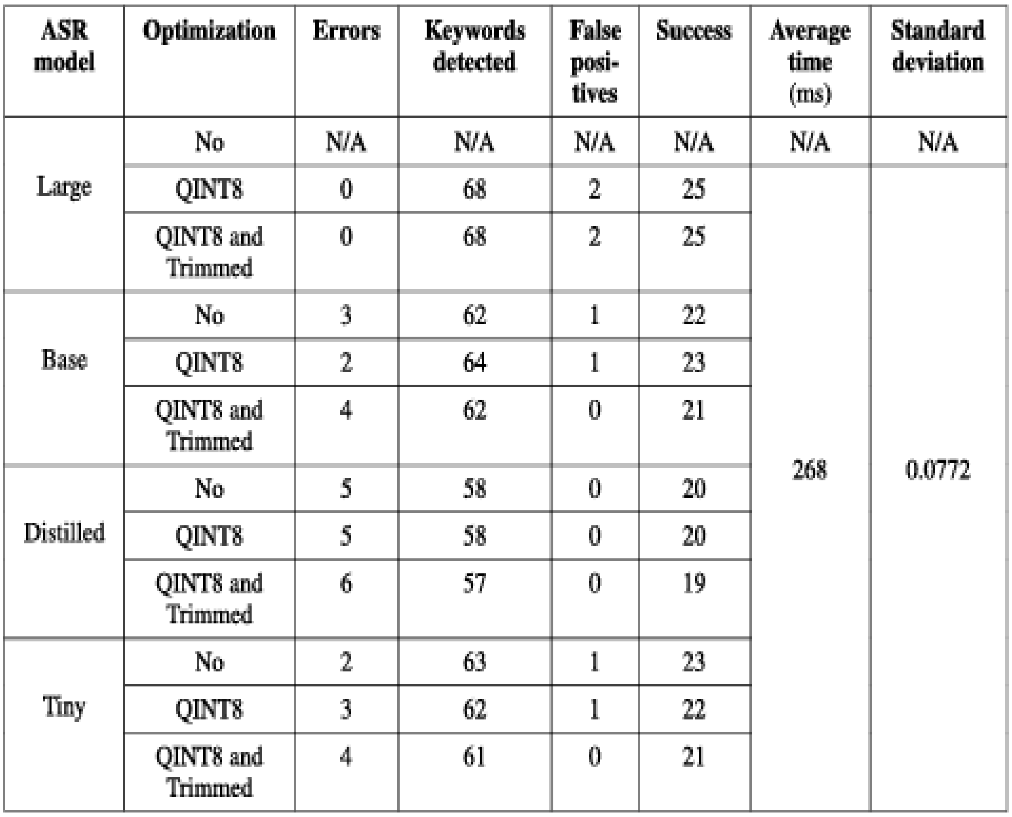
TABLE 4.1 Latency results obtained for transcription inference with large model hosted on the cloud when tested in scenarios A and B.



Scenario	**A**						**B**					
	Query 1		Query 2		Inference	RTT	Query 1		Query 2		Inference	RTT
Time	Connect	Total	Connect	Total			Connect	Total	Connect	Total		
Average (ms)	40.81	176.13	39.95	136.07	336.5	41.27	101.77	1301.57	275.62	614.16	1938.3	56.38
Standard deviation	2.33	10.27	1.4	7.9	10.95	2.04	195.67	1787.31	576.6	1485.73	3263.79	20.459
Minimum (ms)	39.33	169.82	39.16	127.25	323	39	44.59	602.42	58.38	175.21	817	44.15
Maximum (ms)	47.68	21.51	44.16	156.58	370	50.73	914.35	8813.27	2687.97	6897.31	15735	143.68

This table 4.1 presents the latency results obtained for transcription inference using a large model hosted on the cloud, specifically tested in two scenarios labeled as A and B. Latency refers to the time taken for the system to process and respond to input, indicating its responsiveness and efficiency in transcription tasks.

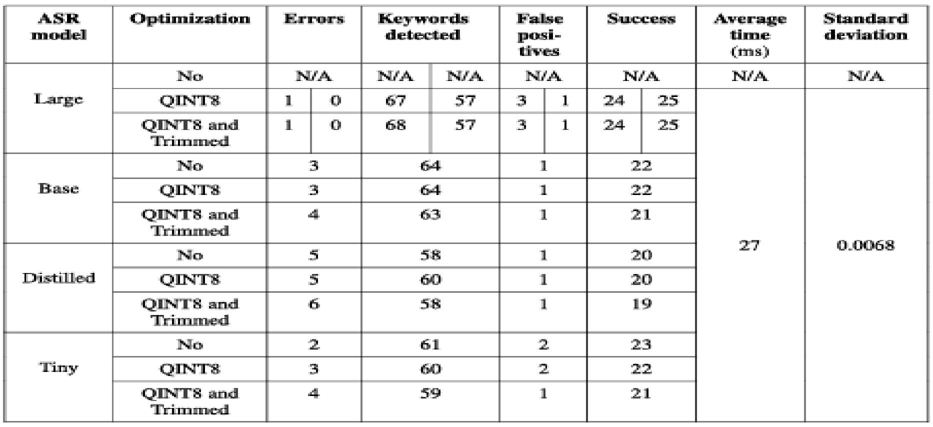
TABLE 4.2 Command detection results obtained with Levenshtein algorithm.



ASR model	Optimization	Errors	Keywords detected	False posi-tives	Success	Average time (ms)	Standard deviation
Large	No	N/A	N/A	N/A	N/A	N/A	N/A
	QINT8	0	68	2	25	268	0.0772
	QINT8 and Trimmed	0	68	2	25		
Base	No	3	62	1	22		
	QINT8	2	64	1	23		
	QINT8 and Trimmed	4	62	0	21		
Distilled	No	5	58	0	20		
	QINT8	5	58	0	20		
	QINT8 and Trimmed	6	57	0	19		
Tiny	No	2	63	1	23		
	QINT8	3	62	1	22		
	QINT8 and Trimmed	4	61	0	21		

This table 4.2 presents the command detection results obtained using the Levenshtein algorithm, a method commonly employed for measuring the similarity between two sequences. The table likely provides insights into the effectiveness of the algorithm in accurately detecting commands within speech utterances.

TABLE 4.3 Command detection results obtained with LCS algorithm.



ASR model	Optimization	Errors	Keywords detected	False posi-tives	Success	Average time (ms)	Standard deviation
Large	No	N/A	N/A N/A	N/A	N/A	N/A	N/A
	QINT8	1 0	67 57	3 1	24 25	27	0.0068
	QINT8 and Trimmed	1 0	68 57	3 1	24 25		
Base	No	3	64	1	22		
	QINT8	3	64	1	22		
	QINT8 and Trimmed	4	63	1	21		
Distilled	No	5	58	1	20		
	QINT8	5	60	1	20		
	QINT8 and Trimmed	6	58	1	19		
Tiny	No	2	61	2	23		
	QINT8	3	60	2	22		
	QINT8 and Trimmed	4	59	1	21		

This table 4.3 provides an analysis of command detection results obtained using the Longest Common Subsequence (LCS) algorithm. The LCS algorithm is commonly used for comparing sequences of elements to find the longest subsequence shared by both sequences, making it suitable for detecting similarities between spoken commands and predefined command sequences.

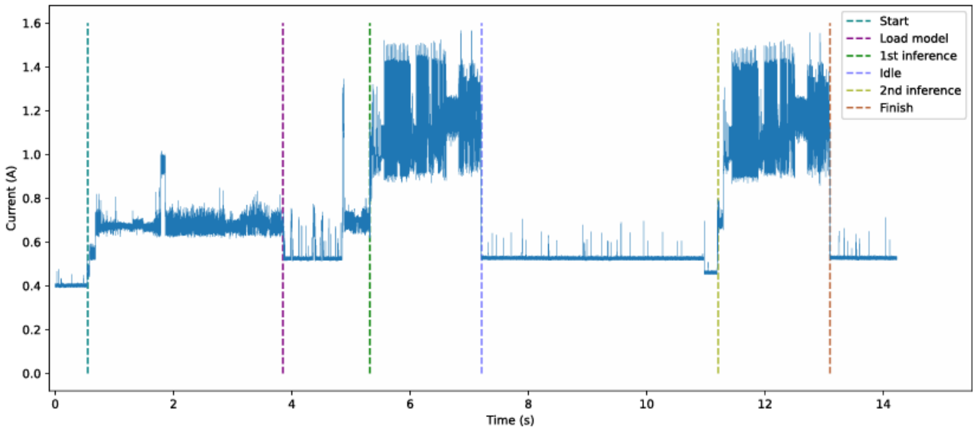
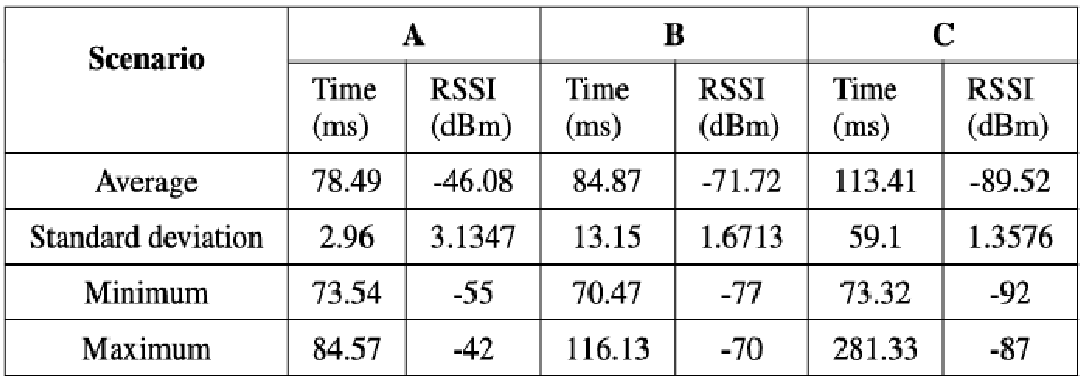


FIGURE 4.1 Regular operation of the Distilled model with 2 speech inferences.

4.2 DISCUSSION ON RESULTS

In this section, the paper delves into a detailed discussion and analysis of the obtained results from the experiments or evaluations conducted. The discussion aims to provide insights, interpretations, and contextualization of the findings, shedding light on the implications and significance of the results for the study or project at hand.

TABLE 4.3 Results obtained for the Bluetooth communications subsystem when testing and executing IoT commands in the three selected scenarios.



Scenario	A		B		C	
	Time (ms)	RSSI (dBm)	Time (ms)	RSSI (dBm)	Time (ms)	RSSI (dBm)
Average	78.49	-46.08	84.87	-71.72	113.41	-89.52
Standard deviation	2.96	3.1347	13.15	1.6713	59.1	1.3576
Minimum	73.54	-55	70.47	-77	73.32	-92
Maximum	84.57	-42	116.13	-70	281.33	-87

The results shown in Table 4.3 indicate that, on aver age, in the scenario with the worst signal (C) a latency of 113.41ms was required, while only 78.49ms for the best signal scenario(A).In any case, less than 282ms were necessary.

These results show that BLE Mesh is an efficient proto col for performing communications with IoT devices when short messages are involved and when considering the total time as the time of transcription and command detection. For instance, in an Android-based system it would be possible with the fastest models to perform an action in less than 1 second, which is a reasonable latency for immediate-response tasks such as lighting control. For completion purposes, the next section goes further and analyzes latency for opportunistic scenarios, where no direct communications are available.

1) VAD OPPORTUNISTIC COMMUNICATIONS

OPPORTUNISTIC SCENARIO

In opportunistic communications, mobile devices communicate directly when in range, beneficial in urban areas or where no infrastructure exists. In smart homes, traditional systems use centralized hubs, while VADs act as mobile nodes, receiving messages opportunistically. This contrasts with static IoT devices. Replicating information between relay nodes enables VADs to function without centralized hubs, reducing infrastructure costs.

OPPORTUNISTIC PERFORMANCE

To assess the impact of opportunistic message strategy, the delay in redirecting messages through intermediate gateways was measured. Total system latency isn't crucial because LPN nodes are often asleep due to power constraints, and the information they send to the VAD node isn't time-sensitive. For example, in activating the intelligent heating system, rapid temperature polling isn't necessary. System response latency depends on LPN node sleep cycles. However, the time to circulate an end-to-end message to the VAD node across different hops was measured, ignoring sleep times. Table displays transmission times for varying numbers of hops in different scenarios, following similar RSSI ranges as detailed earlier. Testing also considers the increase in intermediate nodes to ensure messages aren't skipped. Lowering transmission power levels of nodes helps avoid skipping intermediate hops in BLE Mesh communication.

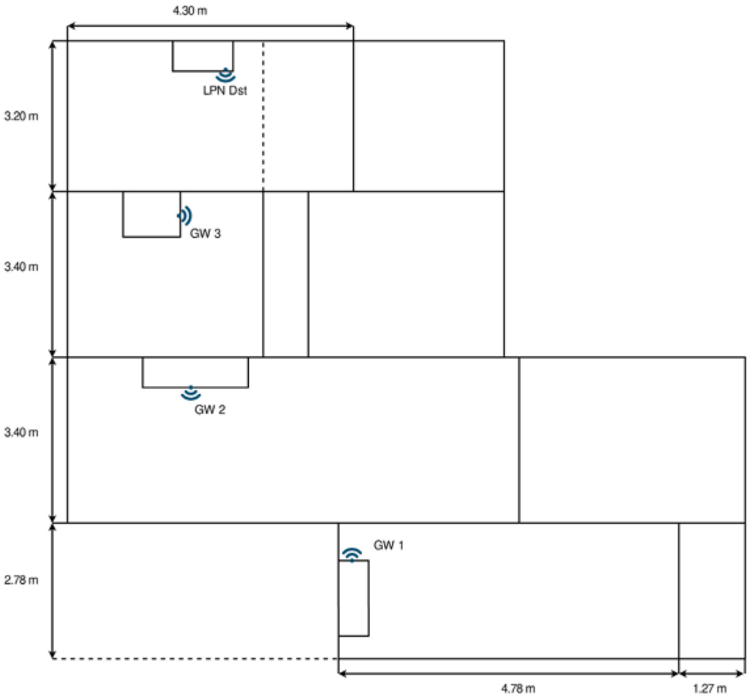
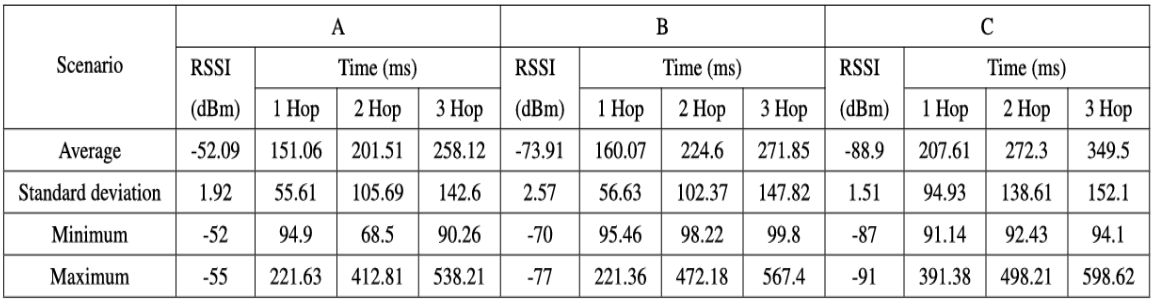


FIGURE 4.2 Node and gateway locations for the indoor tests.

The figure 4.1 represents the tests were carried out indoors in a house and through out 8 different rooms. A map of the indoor scenario in which the tests were carried out.In such a Figure, GW1, GW2 and GW3 represent the intermediate gateways employed in the different scenarios (the gateway antenna orientation is also depicted). The destination node

(LPN Dst) remained for all tests in the same position, while the intermediate gateway position varied according to the number of hops: for one hop only GW2 was used; with two hops, GW2 and GW1 were employed; and for three hops all the represented gateways were used. In addition, the inter mediate gateways of the different scenarios were static, with stable RSSI values that ranged between-66 and-82dBm. In contrast, the VAD node was placed in different positions depending on the scenario and considering the RSSI ranges defined in Section IV-G.

TABLE 4.4 Latency results obtained in multi-hop communication between the VAD and the IoT node.



Scenario	A				B				C			
	RSSI	Time (ms)			RSSI	Time (ms)			RSSI	Time (ms)		
	(dBm)	1 Hop	2 Hop	3 Hop	(dBm)	1 Hop	2 Hop	3 Hop	(dBm)	1 Hop	2 Hop	3 Hop
Average	-52.09	151.06	201.51	258.12	-73.91	160.07	224.6	271.85	-88.9	207.61	272.3	349.5
Standard deviation	1.92	55.61	105.69	142.6	2.57	56.63	102.37	147.82	1.51	94.93	138.61	152.1
Minimum	-52	94.9	68.5	90.26	-70	95.46	98.22	99.8	-87	91.14	92.43	94.1
Maximum	-55	221.63	412.81	538.21	-77	221.36	472.18	567.4	-91	391.38	498.21	598.62

Table 4.4 presents latency results from multi-hop communication between the VAD (Voice Assistant Device) and the IoT (Internet of Things) node. The table likely includes data on transmission times for messages traveling through different numbers of intermediate nodes (hops) in various scenarios. These scenarios may involve opportunistic communication setups, reflecting different RSSI (Received Signal Strength Indication) ranges outlined in Section IV-G of the document. The table helps assess the performance of the system in terms of message transmission delays under different conditions.

CONCLUSION

In this paper, an IoT home automation system with voice assistance has been presented. The system operates exclusively on the edge, without the need for an Internet connection, requiring a reduced deployment in infrastructure and allowing the use of low-resource languages like Galician. The system has been tested in static and mobile opportunistic scenarios, providing relatively fast and efficient response times, performing CPU-only inference for transcriptions, being suitable for edge devices with minimal computational capabilities and implementing a distributed architecture without the need for expensive gateways or hubs to manage the communications with the deployed IoT nodes.

Specifically, different multi-language ASR models were Developed, validated and optimized for being used with a voice assistant. A relevant effort was put on the optimization and performance evaluation of reduced models that can be used by Android devices or by other types of resource-constrained embedded hardware. Moreover, the proposed system considers the use of mobile opportunistic devices, which is beneficial from the infrastructure point of view. Such an opportunistic approach was evaluated in terms of latency, obtaining the system response times in different scenarios. In terms of latency, with the most limited hardware, inferences of less than 2 s were achieved for the best case while accuracy for the worst case reached a success rate of 76%. Specific details have been provided regarding the energy consumption required by the transcriptions of each tested model. Finally, transmission latency was measured between the VAD and an IoT nodes, showing that it never exceeded 300 ms for the worst case in direct connection and 600 ms with 3 hops.

REFERENCES

[1] A. Ghosh, D. Chakraborty, and A. Law, “Artificial intelligence in Internet of Things,” CAAI Trans. Intell. Technol., vol. 3, no. 4, pp. 208–218, Dec. 2018.

[2] P. Kaur, P. Singh, and V. Garg, “Speech recognition system; challenges and techniques,” Int. J. Comput. Sci. Inf. Technol., vol. 3, no. 3, pp. 3989–3992, 2012.

[3] S. F. N. Zaidi, V. K. Shukla, V. P. Mishra, and B. Singh, “Redefining home automation through voice recognition system,” in Emerging Technologies in Data Mining and Information Security (Advances in Intelligent Systems and Computing), vol. 1300, A. E. Hassanien, S. Bhattacharyya, S. Chakrabati, A. Bhattacharya, and S. Dutta, Eds. Singapore: Springer, 2021, doi: 10.1007/978-981-33-4367-2_16.

[4] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, “Edge intelligence: Paving the last mile of artificial intelligence with edge computing,” Proc.IEEE, vol. 107, no. 8, pp. 1738–1762, Aug. 2019.

[5] Voice Recognition Tech Privacy and Cybersecurity Concerns. Accessed: Feb. 2023. [Online]. Available: https://www.natlawreview.com/article/voice-recognition-technology-market-surges-organizationsface-privacy-and

[6] J. Lau, B. Zimmerman, and F. Schaub, “Alexa, are you listening?” Proc.ACM Hum.-Comput. Interact., vol. 2, pp. 1–31, Nov. 2018.

[7] A.-L. Georgescu, A. Pappalardo, H. Cucu, and M. Blott, “Performance vs. hardware requirements in state-of-the-art automatic speech recognition,” EURASIP J. Audio, Speech, Music Process., vol. 2021, no. 1, pp. 1–30, Jul. 2021.

[8] S. Gondi and V. Pratap, “Performance evaluation of offline speech recognition on edge devices,” Electronics, vol. 10, no. 21, p. 2697, Nov. 2021.

[9] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, “Applying wav2vec2.0 to speech recognition in various low-resource languages,” 2020, arXiv:2012.12121.

[10] E. Guglielmi, G. Rosa, S. Scalabrino, G. Bavota, and R. Oliveto, “Sorry,I don’t understand: Improving voice user interface testing,” in Proc. 37th IEEE/ACM Int. Conf. Automated Softw. Eng., Oct. 2022, pp. 1–12.

[11] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 12449–12460.

[12] Language Diversity Index. Accessed: Feb. 2023. [Online]. Available:https://education.nationalgeographic.org/resource/language/diversityindex-map

[13] Ageing Europe—Statistics on Population Developments. Accessed:Feb. 2023. [Online].Available:https://ec.europa.eu/eurostat/statistics/explained/index.php?title=Ageing_Europe__statistics_on_population_developments

[14] Instituto Galego de Estatística. Accessed: Feb. 2023. [Online]. Available:https://www.ige.gal/web/index.jsp

[15] F. N. Valverde and M. Labianca, “Depopulation and aging in rural areas in the European Union: Practices starting from the LEADER approach,”Perspect. Rural Develop., vol. 2019, no. 3, pp. 223–252, 2019.

[16] I. B. C. Irugalbandara, A. S. M. Naseem, M. S. H. Perera, and V. Logeeshan, “HomeIO: Offline smart home automation system with automatic speech recognition and household power usage tracking,” in Proc. IEEE World AI IoT Congr. (AIIoT), Jun. 2022, pp. 571–577.

[17] N. Chumuang, M. Ketcham, S. Tangwannawit, W. Yimyam, S. Hiranchan, M. Rattanasiriwongwut, and P. Pramkeaw, “Development a home electrical equipment

control device via voice commands for elderly assistance,” in Proc. 15th Int. Joint Symp. Artif. Intell. Natural Lang. Process. (iSAINLP), Nov. 2020, pp. 1–7.

[18] L. Xu, A. Iyengar, and W. Shi, “CHA: A caching framework for homebased voice assistant systems,” in Proc. IEEE/ACM Symp. Edge Comput.(SEC), Nov. 2020, pp. 293–306.

[19] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, “XLS-R: Self-supervised cross-lingual speech representation learning at scale,” in Proc. Interspeech, Sep. 2022, pp. 2278–2282.

[20] SparkFun Edge Development Board—Apollo3 Blue MCU—DEV-15170—SparkFun Electronics. Accessed: Feb. 2023. [Online]. Available: https://www.sparkfun.com/products/15170.