Transforming Daily Tasks for Visually Impaired Seniors: Leveraging an OpenAI Model-Enhanced Multimodal Dialogue System with Speech Recognition in POF Smart Homes

Jason Zheng[1], Paloma Almonte[1], Jackson Haft[1], Dayana Ribas[2], Alfonso Ortega[2], Eduardo Lleida[2], Pedro Luis Carro[2], María Ángeles Losada[2], Alicia López[2], Javier Mateo Gascón[2], Yi Wang[1], Neo Antoniades[3], Dwight Richards[3], Jessica Jiang[3]

[1]Electrical and Computer Engineering Department, Manhattan College, Riverdale, New York, USA 10471

[2]Electronic and Communications Engineering Department, Aragón Institute of Engineering Research, University of Zaragoza, Spain 50018

[3]Engineering & Environmental Sciences Department, College of Staten Island, The City University of New York, New York, USA 10314

***Abstract* – We proposed a Client Server Based Plastic Fiber Optic Network for an Elderly Support system in a smart home environment. The network will incorporate multimodal dialogue systems (MDS) and Artificial Intelligence (AI) systems. Existing AI models like GPT and Vicuna will be used and further trained to support the elderly in a smart home. The MDS systems, Whisper and Lavis, will act as intermediaries between the POF network of sensors and the AI systems. The systems will convert video and audio files into information that the AI systems can process and respond to. The POF network, consisting of sensors and a client-server architecture, will serve as the system's backbone. Its primary purpose is to gather data for the AI system and act based on its output. This research aims to enhance the safety, well-being, and independence of the elderly by leveraging advanced network technologies.**

Keywords—Plastic fiber optic network, Elderly support system, Client-server architecture, Artificial Intelligence (AI), Multimodal Dialogue System (MDS)

I. INTRODUCTION

The aging population presents unique challenges and demands for healthcare and support systems. There is still a lack of research tailored to the unique requirements and limitations of elderly individuals, particularly those with visual or hearing impairments. In conjunction to this challenge, there has been significant traction within the fiber optic space, specifically around Plastic Fiber Optics (POF). This traction was created from the commercialization of POF through the term, "FTTX", which has been generated in industry to imply, "Fiber to the x", where x can represent home, node, curb or building. The research will allow visually impaired elderly individuals to live comfortably and safely in their own homes for as long as possible. This can help reduce the burden on healthcare systems and improve the overall care for the aging population.

This research aims to address these challenges by developing a smart Client-Server Based Plastic Fiber Optic Network for an Elderly Support System. The objective is to establish a robust and low-latency fiber optic network infrastructure in the homes of elderly individuals, bringing smart automation and real-time data transmission. The network infrastructure will take advantage of neural networks, natural language processing, and a multimodal dialogue system to assist the user in daily tasks and more unique and uncommon requests. Using a combination of sensors, microphones, speakers, and cameras, the neural network will make real-time updates and make suggestions to the user based on requests, stages of the day, or current events that are happening around the user. Such AI systems will take advantage of existing models such as GPT[1] and Vicuna [2]. We plan on programming them, so they are better suited to fill in the role of a neural network for an elderly smart home.

The research will explore a multimodal dialogue system that integrates Faster Whisper [3], a state-of-the-art voice-to-text conversion tool [4], and Lavis [5], an advanced object recognition system [6]. This study is particularly aimed at enhancing the accessibility and usability of Natural Language Processing (NLP) applications for elderly individuals, especially those with visual or hearing impairments. The key focus is on utilizing Whisper to accurately convert the speech of elderly users into text and employing Lavis to provide real time object descriptions through a camera interface. The research objectives are twofold: firstly, to evaluate the accuracy and effectiveness of Whisper in transcribing elderly speech into text, ensuring that responses are accurately captured and conveyed. Secondly, the study aims to assess the usability and user experience of Lavis in delivering precise and helpful object descriptions. Additionally, the research will explore the combined impact of Whisper and Lavis on improving the interactions between elderly individuals and NLP applications, potentially revolutionizing the way this demographic engages with technology. Both Whisper and Lavis are freely available models, underscoring their accessibility and potential for widespread application in this field.

By setting specific tasks for the system in mind, we plan to integrate the LLM to only cover some initial tasks at first. The three tasks include being able to search for local weather, searching for recipes based on what the user requests, and

being able to update a user calendar. With these tasks in mind, we try to find the best way for LLMs to complete said task while being in a whole system. These tasks are just the baseline, and the end goal is to have a system that is capable of learning and making changes to itself to suit the user better [19-20].

The integration of MDS tools within the proposed smart home system tailored for the elderly population raised pertinent inquiries pertaining to the physical interface aspect of the prototype. The mechanism of converting speech into text is accomplished by utilizing the 2022 OpenAI Whisper model. An essential investigation encompassed evaluating the accuracy and dependability of Whisper in transcribing speech from elderly individuals into text and text-to-speech, alongside the identification of potential limitations linked to speech patterns, accent variations, or speech impairments unique to this demographic.

While the MDS models offer substantial advantages, it is imperative to comprehend the associated implementation costs when incorporating the model into the broader system architecture, as well as to determine the most suitable model for integration. Throughout the research endeavor, we systematically delved into critical inquiries to advance our comprehension of the smart home system's performance dynamics. Of notable significance was an examination of the temporal delay users might experience before receiving responses or actions from the system. Subsequently, a probing investigation sought to gauge the feasibility of optimizing response times to ensure seamless and efficient interactions with the system. The addressing of these fundamental questions serves as a pivotal cornerstone for a comprehensive evaluation of the system's user experience, thereby gauging its efficacy in meeting user expectations.

The original OpenAI Whisper framework encompasses five models—tiny, base, small, medium, and large. Furthermore, our analysis encompassed a study of various enhanced Whisper models, including Whisper CPP and Faster Whisper. Across our research initiatives, the Faster Whisper medium model consistently demonstrated superior performance in diverse tests, effectively striking a harmonious equilibrium between accuracy and inference time for the system.

Main contributions and novelty components: this study aims to produce new/novel information by developing a comprehensive Client-Server Based Plastic Fiber Optic Network for an Elderly Support System. This system will be novel in the fact that it will actively train and optimize the network's environments based on the user's requests across the time of its time of operation. The system will also have a hierarchical structure to its commands as voice identification will allow the network to decisively know who its main user is and prioritize their request over other users within the network's environment. Maintaining this oversight will allow the elderly user to be consistently accommodated and prioritized within the comfort of their own home as the network also monitors the environment around them.

II. Related Work

Recent research introduced an innovative automatic health monitoring system for detecting voice disorders, specifically designed for the evolving smart city infrastructure. Utilizing the advancements in Internet of Things (IoT) and cloud computing, this system adeptly addresses the increasing need for efficient healthcare solutions in urban areas with a growing elderly population [18-20]. It employs linear prediction analysis to differentiate between normal and voice-disordered subjects, analyzing voice samples to extract crucial features. The system boasts remarkable accuracy, with 99.94% ± 0.1 for sustained vowels and 99.75% ± 0.8 for running speech, making it a highly effective tool for early detection and monitoring of voice complications across diverse age groups and professional backgrounds in smart cities.

Bennett et al [7] talks about the evolution of smart healthcare technologies that are showing in a transformative era for healthcare delivery and the daily life of people. This research paper delves into the rapidly changing and evolving environment of intelligent healthcare solutions, exploring their potential to revolutionize health monitoring, improve wellness, and enhance the quality of life within the comfort of one's home. From wearable sensors that provide continuous and accurate data collection to robotics and personal assistants poised to assist and empower individuals, these innovations are reshaping how healthcare is perceived and practiced. However, while these advancements hold great promise, they also introduce many challenges that need to be addressed. Ethical considerations, security concerns, and questions about the reliability of data generated by these technologies raise important issues that must be carefully navigated.

Sai Kiran et al [8] introduces the "Home Intruder Detection System (HIDES)," an innovative solution aimed at enhancing home security using Internet of Things (IoT) technology. The system utilizes the Single-Shot Multibox Detection (SSD) algorithm implemented on NVIDIA Jetson Nano, enabling the accurate identification of intruders through connected cameras. This advanced algorithm ensures the system's ability to process live video frames swiftly and promptly detect potential threats. Additionally, the system features a minimalist yet efficient mobile application that facilitates remote monitoring and alerts for homeowners, enabling them to take immediate action in response to any detected intrusions.

However, the study also acknowledges several limitations inherent to the current implementation of HIDES. When comparing hardware options, the performance gap between NVIDIA Jetson Nano and Raspberry Pi is highlighted, with the latter experiencing significantly lower frame rates. This aspect could impact the system's ability to provide real-time alerts. Another algorithmic consideration is the sensitivity difference observed between the chosen SSD algorithm and the You Only Look Once (YOLO) algorithm. The latter's lower sensitivity in detecting individuals might affect the system's accuracy in identifying potential intruders. Furthermore, the minimalist design of the mobile application, while contributing to its efficiency, may lack certain advanced features offered by other comprehensive security apps available in the market. The placement of hardware components also poses challenges, as

improper positioning of the detection system or cameras could compromise the system's overall effectiveness. The dependence on uninterrupted internet connectivity and a reliable power source is emphasized, as any disruption in these areas could render the system inactive during crucial moments.

III. Background

3.1. POF

The study builds upon the principles of plastic fiber optic networks integrated with smart automation through a client-server architecture. Plastic fiber optic (POF) cables provide the backbone of the network infrastructure, offering high bandwidth, security, lower operational cost as well as lower latency between hardware connections [21-22]. The high data rate of fiber optics through the implementation of high-speed transceivers, connectors, and ethernet ports allows for faster sending and receiving of data from both the client and the server. Client-server architecture allows the multimode POF to facilitate efficient communication from one client to server as well as allow for a future higher data rate connection as more multi-client connections are created within the network. Integrating the clients with strategically placed sensors, including audio, visual imaging, and motion sensors, allows the server to receive relevant data. Allowing for automation of tasks oriented towards the user while also integrating with your Google account to create and sink information, allowing for the users to create calendar events or search the web via ChatGPT-based server connection.

3.2 AI Model

Using the POF as a backbone, automatic speech, image, and motion recognition systems will work in conjunction with an AI (artificial intelligence) system that relies on natural language processing. Studies have been conducted on this topic before, but due to a lack of sophisticated natural language processing, there has been a limit to what the AI system can do and how effectively the task can be performed [9]. With more advanced natural language processing systems such as GPT and Vicuna, we are able to take in data from the automatic recognition systems and transform them into a response for the user and help them perform any reasonable task.

According to GPT 3.5 and Vicuna, they comprise eight components: Transformer Architecture, Pre-training Data, Reinforcement Learning, NLP, machine learning algorithm, hardware infrastructure, and data management system. Transformer Architecture is a deep learning model architecture specifically designed for processing sequential data, such as natural language. Transformers utilize self-attention mechanisms to capture contextual relationships between words or tokens in a sentence or sequence [10].

Pre-training data is all the data taken by GPT to help the model learn patterns, grammar, facts, and contextual relationships in language [11]. Supervised fine-tuning is training the model based on select datasets and having human AI trainers interact with the model to help shape its behavior and responses [11]. Reinforcement learning involves more fine-tuning through rewarding the model when it generates a desired response and penalizing it when it generates inappropriate responses. These components talk about setting up the AI model and the initial training.

The 4th component, NLP techniques, are used to understand the meaning and context of the input text and generate appropriate responses [12]. This includes techniques such as part-of-speech tagging, named entity recognition, sentiment analysis, and context-aware word embedding. The 5th component, Machine learning algorithms, are used to learn from the vast amount of text data that these systems are built on and to improve their performance on various NLP tasks. This includes algorithms such as deep learning models, neural networks, and gradient descent [12]. Hardware infrastructure is the 7th component and is often overlooked. AI needs to run on powerful hardware infrastructure such as highperformance servers and cloud computing platforms, which allows the model to process large amounts of text data and generate responses in real time [13]. Finally, a data management system is needed to store and retrieve the vast amount of text data that these models train on and store. This includes systems such as distributed databases, data warehouses, and data processing frameworks [14].

Overall, the combination of these components allows to understand and generate human-like text and to perform various NLP tasks with high accuracy and efficiency. This makes it possible for them to fit in a smart home environment since they are trainable and can think to an extent and make choices.

Finally, ethical guidelines are there as safety measures to ensure the responsible and ethical use of the model. These guidelines aim to mitigate issues like biased behavior, inappropriate responses, or the spreading of misinformation [15]. While it might be a component that is not obvious in an AI model, it is something that can be a huge concern. We need to be able to train AI in a way that makes sure it doesn't have a bias toward any group of people. We also need to make sure that it never develops any malicious intent. This might not always be the case as AI develops and becomes able to train itself more and more, and we must have a failsafe [16].

3.3 MDS System

The NPL will be obtaining data and the requested information from the user before making any calculations. The MDS will contain Whisper, Lavis, and Coqui, which are three distinct components that are often used together in the context of speech-to-text conversion and object recognition.

Whisper is an automatic speech recognition (ASR) technology that uses algorithms to analyze and interpret audio input, typically in the form of speech, and transcribes it into a textual form, specifically focused on acoustic modeling [3-5]. It uses deep learning techniques, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), to model the relationship between audio signals and corresponding text transcripts. Whisper incorporates four language models to improve the accuracy of speech recognition. Language models help predict the most likely sequence of words given the context and previous words in a sentence. Whisper leverages transfer learning, which involves pre-training models on large amounts of data and fine-tuning them for specific tasks. This allows the model to learn from a

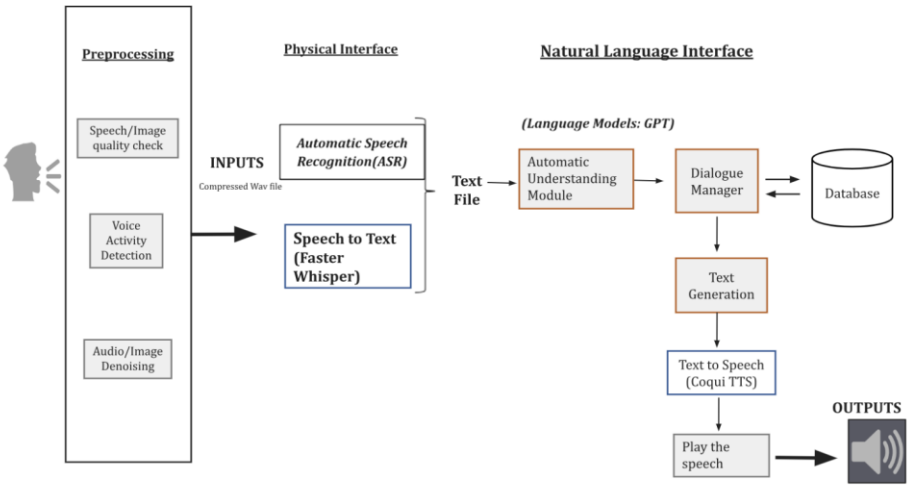

Fig. 1. PoF Smart Home System Design

diverse range of speech data and improve its performance on specific domains or tasks.

Lavis utilizes CNNs for object recognition in images. CNNs are neural networks specifically designed for analyzing visual data, such as images. They employ convolutional layers to extract meaningful features from images and make predictions based on those features. It employs deep learning techniques to extract high-level features from images. Lavis can recognize objects at different levels of abstraction, allowing for accurate object recognition. Lavis performs both object detection and classification. Object detection involves identifying and localizing multiple objects within an image, while object classification focuses on assigning specific labels or categories to those objects.

Coqui is an open-source automatic speech recognition (ASR) toolkit that allows users to convert text to speech. It aims to provide accessible and customizable ASR capabilities for various applications. Coqui is built on the Mozilla DeepSpeech project, which employs deep learning models and the TensorFlow framework for ASR tasks. Coqui enables users to train their own speech recognition models using their own data. It provides tools and resources for data preparation, model training, and optimization, allowing users to customize and improve ASR performance for their specific use cases.

IV. Materials and Methods

4.1 *Background*

The technology we will be investigating consists of both physical and software-based systems. On the physical layer, we utilized two configurations of clients, a Jetson Nano with a 128-core NVIDIA Maxwell™ architecture-based GPU outfitted with a push-button and a USB Microphone. The Raspberry Pi 4 Model B with 4GB of RAM along with a CSICamera and PIR sensor integrated into the board. In the initial testing of the following network, the data transmission will be tested via ethernet connection through a TP-Link 5-Port 1GB switch. In a later phase, POF hardware for this project will utilize 1GB Ethernet-POF Media Connectors allowing to both transmit and receive data across the network.

The software integrated into the system is part of the physical interface and the natural language interface. he system architecture depicted in Figure 1 illustrates a comprehensive smart home system design integrating a Physical Interface with a Natural Language Interface, streamlining interaction through speech and language processing. At the preprocessing stage, speech and image quality checks are conducted alongside voice activity detection and denoising, ensuring clear input for the ASR module.

The physical interface part of the system will be focusing on the ASR and conversion of text to speech. ASR is the technology that converts spoken language into written text, enabling the system to understand and process human speech. In a smart home system, ASR allows users to interact with devices using natural speech, facilitating hands-free control and seamless communication with smart devices for various tasks. The Faster Whisper medium model [17] was the chosen ASR for the research since it developed an accurate and fast text file to the natural language interface to use. The faster whisper medium model code was obtained from the opensource Github website, and then it was customized to be integrated into the smart home system.

Integrating Whisper and Lavis into a seamless system embodies a nuanced approach to enhancing smart home environments, particularly in applications requiring both auditory and visual comprehension. The journey begins with capturing audio through strategically placed microphones, which collect the user's spoken commands. These audio inputs are the domain of Whisper, a sophisticated speech-to-text engine that translates verbal communication into written text. Its capacity to accurately interpret speech across a

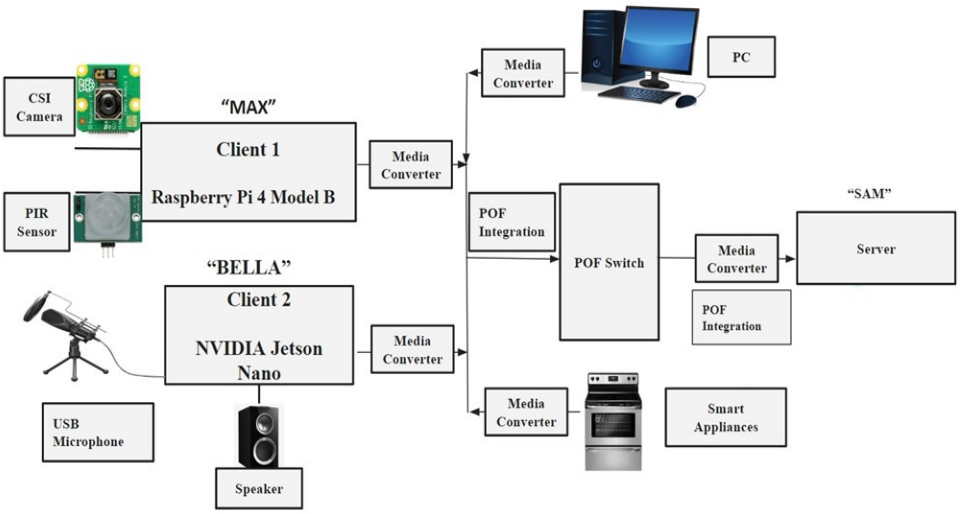

Fig. 2 Hardware Components of the Proposed PoF Smart Home System Design System

myriad of languages, accents, and dialects is crucial, as it forms the foundation upon which the system understands user requests.

Following the transcription process, the system analyses the generated text to discern whether the user's request necessitates visual verification or information. This step is pivotal, as it determines the subsequent engagement of Lavis, an object recognition model. For instance, if a user inquires about the location of a personal item, such as glasses, the system identifies this as a task for Lavis, which then springs into action.

Upon activation, Lavis begins processing visual data from cameras installed within the premises. Its deep learning algorithms sift through the visual input to pinpoint the requested item, demonstrating its prowess in object recognition under varied conditions. This process not only showcases the model's accuracy but also its adaptability to different environments and lighting situations.

The final step in this integrated system involves communicating the outcome to the user. Depending on the setup and user preferences, this feedback can be auditory, through synthesized speech, or visual, via displays. This stage is where the user is informed about the located item's whereabouts, thus completing the cycle of interaction that began with a simple verbal request.

The software behind the Natural Language Interface would consist of OpenAI's GPT-3.5 model. This part of the software side is responsible for processing the text inputs fed to it by the ASR software. Using GPT as the base for the API, we use other readily available APIs such as Google Calendar API and Serpai API to, respectively, let us update a user's calendar or Google search for some information that isn't already available in GPT's database. Tests have also been done with Vicuna due to it being on par with the level of responses given by GPT. Still, ultimately, resource constraints and the simplicity of integrating GPT made GPT the better choice for this test.

The last part of the process will be giving an output to the user, which is achieved by integrating Coqui TT. The Python library provides a natural and customizable speech synthesis solution. When incorporated into a smart home setup, Coqui TTS enhances the system's functionality by enabling real-time conversion of response text files into speech. This integration offers users interactive voice responses, ranging from weather updates to adding events on the calendar, creating a more userfriendly and accessible smart home experience. The process involves the smart home system generating a response text, which is then processed by Coqui TTS using pre-trained models to produce expressive synthesized speech. This speech is relayed to users through speakers or audio devices within the smart home environment, enriching the user interface and catering to diverse user needs, including those with visual impairments or a preference for hands-free interactions. The hardware implementation of the system composes different crucial parts. Figure 2 provides a diagram that serves as a detailed representation of the hardware framework that forms the backbone of a sophisticated smart home automation system. Central to this advanced network are two primary client devices, each with its own specialized function. The first, known as "MAX," is "BELLA," is an NVIDIA Jetson Nano system that interfaces with a USB microphone and speaker, suggesting its utility in processing audio inputs and outputs, possibly for tasks such as voice command recognition or providing audible feedback and communication to the residents.

The integration of these devices into the network involves the conversion of their respective signal outputs — video from the camera, audio from the microphone, and data from the motion sensor provided through their respective ethernet ports and converted via optolock-media converters. This step is indicative of a transformation process, possibly from analog to digital or between digital formats, thus making the data suitable for network transmission.

Both client devices interface with a POF switch, a component of the highlighted POF Integration. PoF is a

technology favored for its short-range communication efficacy, offering notable advantages like high data throughput and resistance to electromagnetic disturbances, attributes highly beneficial for the demands of home automation [21, 22]. The positioning of media converters at both entry and exit points of the POF switch alludes to the adaptation of signals for optical data transfer, ensuring seamless and high-quality communication within the system.

The network's reach extends to a PC and a Server, collectively identified as "SAM." Here, the PC may act as the user interface or control center, while the Server is tasked with more demanding computational workloads, data warehousing, or overall management of the smart home's automation functionalities. The diagram culminates with the inclusion of smart appliances as part of the network, a Raspberry Pi 4 Model B setup, notably equipped with a CSI Camera for visual monitoring and a PIR sensor for detecting motion, indicating its role in security and surveillance within the household. The second, dubbed illustrating the system's broad scope in managing and automating a diverse array of household devices. This integration potentially covers smart refrigerators, ovens, and HVAC systems, among others, all unified under the smart home network to enable sophisticated monitoring and responsive control.

4.2 *Methodology*

Both the ethernet and POF-based connections will be operating on client-server architecture via TCP socket connections. The initiation of the connection will be based on the request of the user, and the server will maintain the connection throughout its interaction with the client allowing for data to be transmitted freely back and forth. On the client side, the GPIO pins were utilized on the Jetson Nano to incorporate the push-to-talk feature. This was done by creating a simple circuit consisting of a 1k resistor in series with the 5V PIN on the jetson nano board. The button was then integrated in parallel with the circuit after the resistor so as the button is pressed, the circuit is broken and the Jetson Nano can detect a low voltage input in turn activating the microphone and beginning the audio acquisition process. Once the button is released, the audio is captured and saved in a ".wav" format, which is then put into a VAD (Voice Activity Detection) that eliminates any non speech or silence within the audio. After this pre-processing, the audio is then saved, the socket connection is opened, and the audio is transmitted to the server. On the server side, once the audio is received, it is a directory called recieved_audio, at which point the newly saved audio is automatically converted into a text file. This text file is then inputted into the running AI which is able to process the request and output a text file which is then automatically converted into an audio file.

V. Experimental Results

5.1 *GPT Results*

We were able to create a preliminary elderly care system based on the premises that we stated previously. The physical layer was created and implemented initially with a Jetson Nano and a laptop-based server connected via Ethernet, which was then upgraded to POF in the network's later stage of development. The Jetson Nano was able to record audio requests from the user by configuring a push-button with the GPIO pins along with a USB-connected microphone. The Jetson Nano operated in a push-to-talk manner allowing the user to record an audio based on the duration of the button being pressed, the audio was then preprocessed directly after through VAD, which allowed for noise and dead space to be removed, resulting in a cleaner audio file to send to the server.

Using some recordings that we captured beforehand, we were able to send them from the server to the client in the form of a ".wav" file. Then using Whisper to transcribe the recordings, we converted the audio file into a text file, which was then fed into GPT-3.5 to process. GPT would then produce a text file stating which had three possible outcomes, which were, stating the task was completed, providing any requested information, or requesting more information to complete the task. Figure 3 is the snippet of code that allows the whole system to run together. The system is demonstrated in Figure 4 by a recipe search through GPT. This was outputted in a text file which would be converted to an audio file using Coqui TTS. This audio file would be sent from the server back to the client in which it would be played out through a connected speaker.

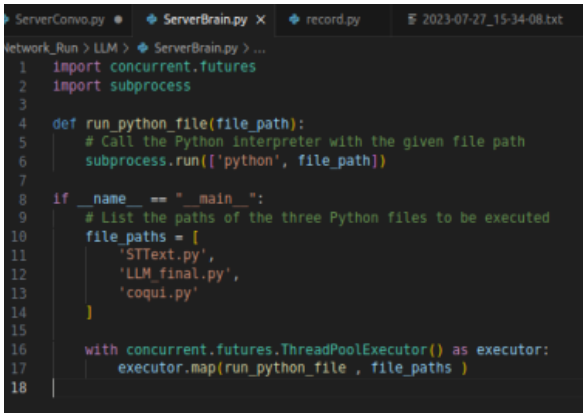

Fig. 3 Code snippet to tie the system together.

5.2 *Whisper Results*

In the context of this research, a pivotal inquiry emerged concerning the Whisper model's dual capabilities of expeditious execution and unwavering precision. To unravel this conundrum, a series of meticulously designed tests were executed. The initial assessment encompassed the determination of the Word Error Rate (WER), a widely employed metric for quantifying the efficacy of an Automatic Speech Recognition (ASR) system. Evaluating the effectiveness of Whisper and Lavis within the smart home system involves several key metrics. Accuracy is paramount, with Whisper's performance measured by its Word Error Rate (WER) and Lavis's efficacy determined by its object identification precision. However, beyond mere accuracy, the user experience—gauged through response times, ease of interaction, and the system's adaptability to unique user needs—forms a critical part of the evaluation. This is particularly relevant when considering the elderly demographic, which may present challenges such as varied

I need to find a recipe with ingredients
Action Input: "chicken and pasta recipe with ingredients"
Observation: Ingredients ; 10 ounces dry pasta (any kind) ; 1 pound boneless skinless chicken breast ; 1 teaspoon garlic powder ; 1 teaspoon Italian seasonin
Thought: I need to find a step by step recipe
Action Input: "step by step chicken and pasta recipe with ingredients"
Observation: Ingredients ; 10 ounces dry pasta (any kind) ; 1 pound boneless skinless chicken breast ; 1 teaspoon garlic powder ; 1 teaspoon Italian seasonin
Thought: I now have the recipe Final Answer:
Ingredients:
10 ounces dry pasta (any kind)
1 pound boneless skinless chicken breast
1 teaspoon garlic powder
1 teaspoon Italian seasoning
1 teaspoon onion powder
1 teaspoon paprika
1/2 teaspoon salt
1/4 teaspoon black pepper
1/4 cup olive oil
2 cloves garlic, minced
1/2 cup chicken broth
1/2 cup heavy cream
1/2 cup grated Parmesan cheese Instructions:
1. Preheat oven to 375°F.
2. Cook the pasta according to package instructions.
3. Meanwhile, season the chicken with garlic powder, Italian seasoning, onion powder, paprika, salt, and pepper.
4. Heat the olive oil in a large skillet over medium-high heat. Add the chicken and cook until golden brown and cooked through, about 5-7 minutes per side.
5. Transfer the chicken to a baking dish.
6. In the same skillet, add the garlic and cook until fragrant, about 1 minute.
7. Add the chicken broth and heavy cream and bring to a simmer.
8. Simmer until the sauce has reduced by half, about 5 minutes. > Finished chain.
'Ingredients: \n10 ounces dry pasta (any kind) \n1 pound boneless skinless chicken breast\n1 teaspoon garlic powder\n1 teaspoon Italian seasoning\n1 teaspoon o cream\n1/2 cup grated Parmesan cheese\nInstructions: \n1. Preheat oven to 375°F. \n2. Cook the pasta according to package instructions.
\n3. Meanwhile, season t the chicken and cook until golden brown and cooked through, about 57 minutes per side. \n5. Transfer the chicken to a baking dish. \n6. In the same skillet, a by half, about 5 minutes.'

Fig. 4 GPT Demonstration

speech patterns due to health conditions, diverse accents, and potential visual impairments that could impact interaction with the system.

The WER is mathematically calculated by SCLITE library; however, to streamline this process, the Python library SCLITE was harnessed. SCLITE, an essential component of the NIST Speech Recognition Scoring Toolkit, facilitates the comparison of recognized speech outputs against reference transcriptions, generating insightful evaluation metrics. Notably, SCLITE automates intricate computations involved in WER calculation. To harness the capabilities of the SCLITE Python library, two indispensable elements are requisitioned: the Ground Truth (reference) and the Hypothesis. The Ground Truth embodies the reference human translation (target) file derived from the test dataset under examination. Contrarily, the hypothesis encapsulates the Machine Translation predictions rendered by the Whisper model for the same test dataset employed in generating the Ground Truth. This dynamic interplay between Ground Truth and Hypothesis lays the foundation for a comprehensive evaluation of Whisper's performance accuracy and speed, rendering a comprehensive portrait of the model's capabilities.

During the testing phase of the Whisper model, a dataset called Thalento, central to the investigation was drawn from the research conducted by the University of Zaragoza [9]. This dataset, comprising approximately 140 audio files, was curated to encompass an equitable distribution of genders and health conditions. Specifically, it included both healthy individuals and those afflicted with pathological conditions, equally distributed between male and female participants. In pursuit of a systematic evaluation, a Python script was meticulously crafted to automate the testing procedure.

The script navigates through discrete folders designated for healthy and pathological subjects, each further subdivided by gender. Operating under the premise of a predefined script, the script scrutinizes the audio contents of each folder, gauging the Whisper model's performance in terms of WER. The average WER for each Whisper model is meticulously computed, shedding light on its efficacy in distinct scenarios. The audio files vary in duration, typically spanning around one minute, contingent on the pace of speech. For comprehensive clarity, Table 1 succinctly encapsulates the average WER percentages, offering a succinct overview of the model's performance across the audio files in consideration.

The large and medium versions exhibited minimal WER percentages, underscoring their remarkable precision in speech recognition. In contrast, the original OpenAI Whisper and the faster iteration showcased marginally lower WERs. Notably, the original faster Whisper boasted an impressive 3.9% WER, a testament to its rapidity in speech analysis. However, the selection of a model for integration mandates a comprehensive outlook. While accuracy remains pivotal, GPU memory usage and processing time assume significance. Remarkably, the faster Whisper model strikes a harmonious balance, optimizing accuracy, GPU memory utilization, and processing time—a crucial facet for real-world integration. It is worth noting that the tiny and base models across all Whisper iterations exhibited relatively higher WERs, leading to their exclusion from further consideration. Thus, while displayed, their original WERs were not factored into subsequent analysis.

However, for this facet of analysis, a nuanced approach was undertaken. The resulting data points were meticulously collated within an Excel spreadsheet, encapsulating the time frames requisite for the completion of audio-to-text conversions across all Whisper model iterations. The culmination of this endeavor is represented in Table 2, which encapsulates the average processing times for each Whisper model. The tiny and base models exhibited notable efficiency, boasting processing times of merely 0.57 and 0.68 seconds, respectively. Conversely, the Whisper CPP medium and large models incurred more substantial temporal overheads, with averages of 37 seconds for the medium variant and 69.92 seconds for the large. The Whisper CPP's substantial processing time renders it an untenable choice.

TABLE I. WISPHER MODLES WORD ERROR RATE

Voice Type	Whisper Models Word Error Rate								
	Tiny	***Base***	***Small***	***Medium***	***Large***	***CPP Medium***	***CPP Large***	***Faster Medium***	***Faster Large***
Pathological	32.9%	23.35%	15.3%	13.9%	9.5%	17.65%	14.84%	12.8%	10.7%
Healthy	23.5%	14.7%	8.5%	5.9%	3.5%	12.35%	9.1%	7.65%	5.92%

TABLE II. WISPHER MODLES AVERAGE TIME TO PROCESS THE THALENO AUDIO DATA

Voice Type	Whisper Models Word Error Rate								
	Tiny	***Base***	***Small***	***Medium***	***Large***	***CPP Medium***	***CPP Large***	***Faster Medium***	***Faster Large***
Average times (s)	.57	.68	1.19	5.44	9.46	37.32	69.92	1.52	2.20

TABLE III. WISPHER MODLES AVERAGE TIME TO LOAD THE THALENO AUDIO DATA

Voice Type	Whisper Models Word Error Rate								
	Tiny	***Base***	***Small***	***Medium***	***Large***	***CPP Medium***	***CPP Large***	***Faster Medium***	***Faster Large***
Average times (s)	2.08	2.09	3.11	5.67	9.75	20.25	32.7	1.57	2.09

for integration within the envisaged smart home system, which necessitates expeditious response times.

Contrastingly, the Faster Whisper medium and large models emerge as the standout contenders, with processing times of 1.52 and 2.20 seconds, respectively. This is owed to their adept utilization of accelerated hardware, which substantially enhances processing speed. In the context of a smart home system, speed is important to ensure seamless user interactions. The Faster Whisper models align harmoniously with the requisites of real-time responsiveness, rendering them the preferred choices for integration.

Table 3 presents a concise overview of model loading times within the smart home system, illustrating the Whisper Medium model's superior speed. The Whisper Medium model achieves an impressive loading time of just 1.57 seconds, outperforming alternative models that require a minimum of 2 seconds for initialization. This stark contrast in loading times, as evident in the tabulated data, underscores the practical significance of the Whisper Medium model's efficiency, promising enhanced user experiences, real-time responsiveness, and optimized system performance.

In Figure 5 illustrates a comparative analysis of the time taken by different Whisper models to process audio files using CPU and GPU resources. This figure is crucial for understanding the computational efficiency of each model in both processing environments. The data shows that processing times on the GPU are significantly shorter due to its parallel processing capabilities, which are better suited for the operations used in machine learning models such as Whisper.

Figure 6 is a graphical representation showing the performance of four different models tested across four trials with varying audio file sizes. The key takeaway from this figure is the consistent performance of the Faster Whisper medium model, depicted in green, indicating its superior speed in processing audio data irrespective of file size. This suggests that the Faster Whisper medium model is highly optimized for speed, making it an ideal choice for real-time applications in smart home environments where quick processing of audio input is critical for user experience and system responsiveness.

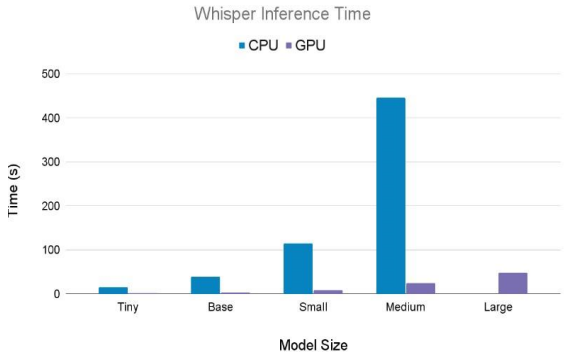

Fig. 5 CPU & GPU ASR Process Time

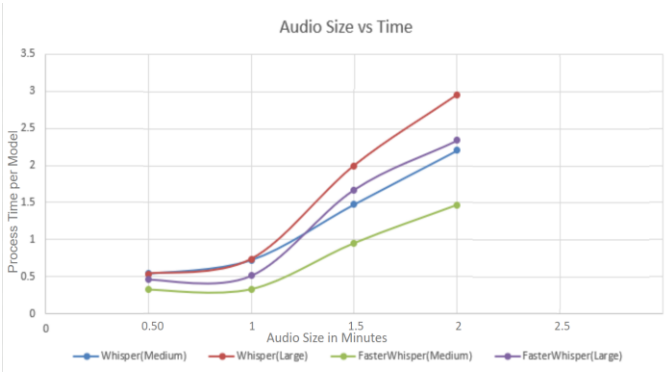

Fig. 6 Process Time Dependability on Audio Size

Figure 6 is a graphical representation showing the performance of four different models tested across four trials with varying audio file sizes. The key takeaway from this figure is the consistent performance of the Faster Whisper medium model, depicted in green, indicating its superior speed in processing audio data irrespective of file size. This suggests that the Faster Whisper medium model is highly optimized for speed, making it an ideal choice for real-time applications in smart home environments where quick processing of audio input is critical for user experience and system responsiveness.

Finally, the implementation of cameras and integration of Lavis into the system is also of great importance. As stated originally, this system was meant to help visually impaired people find their way around their homes and identify objects they might have difficulty locating. Cameras can be activated by a trigger word or by the LLM, which will take a picture of each room and pass it through Lavis, which can then identify which room said requested items are in and pass that information on to the LLM. From there, the system will follow what has already been implemented.

VI. CONCLUSION

In this implementation, we demonstrated the feasibility of a Client-Server Based Plastic Fiber Optic Network for an Elderly Support system in a smart home environment. By setting up a hardware system that is capable of picking up user audio input and outputting audio responses, combined with MDS software and an LLM brain, we were able to create a working system. With modifications and improvements like increased sensors on clients, voice recognition, and localbased storage, the system will be able to adapt to various users while also maintaining the priority of the elders. These modifications can lead to easier patient monitoring via increased sensors, audio inputs/outputs, and cameras while also functioning at a low-energy model through PIR sensors and subfunctions. While the current model shows the basic functions of how the system will operate, with further research, this network can lead to a hyper-customizable fast, paced smart-home environment, being more than capable for providing the elderly with an in-home support system that prioritizes its key user while adhering to various users' live. With the preliminary system set up, we know an elderly home care system is possible using LLM as the brain and POF as the backbone.

However, we do not know how far such a system can go in replacing the need for outside care or a retirement home. We have plans to increase the number of base functions that the system can handle outside of the three already implemented in this experiment. Alongside functions that are hard-coded into the system, we want to be able to have the system identify as well as learn based on daily user inputs and needs allowing for the network to adapt to each user's interactions while always monitoring the key users such as the elderly.

Alongside the software upgrades and development, we hope to also implement safety protocols due to some sensitive information that might be handled by the system. This includes encoding data before being stored in a local or cloud database and so identity confirmation to ensure that the user trying to access the system is authorized. Another section that should be considered is the selection of the most functional VAD library along with other pre-processes, such as noise filtering and keyword detection, making the system run faster and smoother. With the planned work on software and security, there is also an emphasis on the hardware. Picking out suitable devices for the various tasks while falling into a reasonable range is equally important since we don't want such a system to become overly pricey. There is also a need to test which POF material is best suited for this type of system since bandwidth and durability are key factors for the physical implementation of the system.

Another factor that needs to be considered in the future is limiting the amount of GPU and CPU needed to operate the system is of the utmost importance due to both of these being pricey in the long run. On this end, testing with different LLMs, both local and on the cloud, will be tested to see which one is most efficient and requires the least amount of computing power. We need a base system that is both reliable and efficient but, at the same time, cost-efficient. Testing different LLMs will allow us to determine the best efficiency and cost trade-offs.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation Program Standard Grant #2153667, #2153668, under the International Research Experiences for Students (IRES), by the Government of Aragon (T20-23R, T31-23R and T36-23R), by MCIN/AEI/10.13039/501100011033 and the European Union “NextGenerationEU”/PRTR, under grant numbers: PID2021-126061OB-C44, PID2021-122505OBC33 and PDC2021-120846-C4, and by the European Union’s Horizon 2020 research and innovation programme under Marie Skłodowska-Curie Grant 101007666.

REFERENCES

[1] Open AI Documentation, accessed at: https://platform.openai.com/docs/models/gpt-3-5
[2] Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, https://lmsys.org/blog/2023-03-30-vicuna/
[3] Whisper general-purpose speech recognition model. https://github.com/openai/whisper
[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," 2022. [Online]. Available: arXiv:2212.04356,

https://arxiv.org/pdf/2212.04356.pdf

[5] LAVIS - A Library for Language-Vision Intelligence, https://github.com/salesforce/LAVIS)

[6] D. Li, J. Li, H. Le, G. Wang, S. Savarese, and S. C. H. Hoi, "LAVIS: A Library for Language-Vision Intelligence," arXiv preprint arXiv:2209.09019, 2022. [Online]. Available: https://arxiv.org/pdf/2209.09019.pdf

[7] Bennett, J.; Rokas, O.; Chen, L. Healthcare in the Smart Home: A Study of Past, Present and Future. *Sustainability* **2017**, *9*, 840. https://doi.org/10.3390/su9050840J.

[8] K.V.V.N.L. Sai Kiran, R.N. Kamakshi Devisetty, N. Pavan Kalyan, K. Mukundini, R. Karthi, Building a Intrusion Detection System for IoT Environment using Machine Learning Techniques, Procedia Computer Science, Volume 171, 2020, Pages 2372-2379, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.04.257

[9] THALENTO dataset, University of Zaragoza, Spain. https://dihana.cps.unizar.es/~thalento/

[10] Hao, Y., Dong, L., Wei, F., & Xu, K. (2021). Self-Attention Attribution: Interpreting Information Interactions Inside Transformer. Proceedings of the AAAI Conference on Artificial Intelligence, 35(14), 12963-12971. https://doi.org/10.1609/aaai.v35i14.17533

[11] M. Zhang, J. Li, A commentary of GPT-3 in MIT Technology Review 2021, Fundamental Research, Volume 1, Issue 6, 2021, Pages 831-833, ISSN 2667-3258, https://doi.org/10.1016/j.fmre.2021.11.011

[12] R. M. Samant, M. R. Bachute, S. Gite and K. Kotecha, "Framework for Deep Learning-Based Language Models Using Multi-Task Learning in Natural Language Understanding: A Systematic Literature Review and Future Directions," in IEEE Access, vol. 10, pp. 17078-17097, 2022, doi: 10.1109/ACCESS.2022.3149798.

[13] T. Sipola, J. Alatalo, T. Kokkonen and M. Rantonen, "Artificial Intelligence in the IoT Era: A Review of Edge AI Hardware and Software," 2022 31st Conference of Open Innovations Association (FRUCT), Helsinki, Finland, 2022, pp. 320-331, doi: 10.23919/FRUCT54823.2022.9770931.

[14] G. Li, X. Zhou, and L. Cao, AI Meets Database: AI4DB and DB4AI. In Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21). Association for Computing Machinery, 2021, New York, NY, USA, 2859–2866. https://doi.org/10.1145/3448016.3457542

[15] Telkamp, J.B., Anderson, M.H. The Implications of Diverse Human Moral Foundations for Assessing the Ethicality of Artificial Intelligence. J Bus Ethics 178, 961–976 (2022). https://doi.org/10.1007/s10551-022-05057-6

[16] GPT-4 System Card, accessed at https://cdn.openai.com/papers/gpt-4-system-card.pdf

[17] Faster Whisper transcription with CTranslate2, accessed at https://github.com/SYSTRAN/faster-whisper

[18] Almotairi, Khaled H. "Application of internet of things in healthcare domain." Journal of Umm Al-Qura University for Engineering and Architecture 14.1 (2023): 1-12.

[19] Spachos, Petros, Stefano Gregori, and M. Jamal Deen. "Voice activated IoT devices for healthcare: Design challenges and emerging applications." IEEE Transactions on Circuits and Systems II: Express Briefs 69.7 (2022): 3101-3107.

[20] Ali, Z., Muhammad, G., & Alhamid, M. F. (2017). An automatic health monitoring system for patients suffering from voice complications in smart cities. IEEE Access, 5, 3900-3908

[21] Manea, H. K., Molood, Y. N., Al-Jubouri, Q., Taha, B. A., Chaudhary, V., Rustagi, S., & Arsad, N. (2023). A Comparative Study of Plastic and Glass Optical Fibers for Reliable Home Networking. ECS Journal of Solid State Science and Technology, 12(5), 057003.

[22] Wang, S., Liu, B., Wang, Y. L., Hu, Y., Liu, J., He, X. D., ... & Wu, Q. (2023). Machine Learning-Based Human Motion Recognition via Wearable plastic Fiber Sensing System. IEEE Internet of Things Journal.