



**EMBODIED CONVERSATIONAL AGENT WITH
FACIAL EXPRESSIONS**

Munya ALKHALIFA

**2021
MASTER THESIS
COMPUTER ENGINEERING**

**Thesis Advisor
Assist.Prof.Dr. Kasım ÖZACAR**

EMBODIED CONVERSATIONAL AGENT WITH FACIAL EXPRESSIONS

Munya ALKHALIFA

**T.C.
Karabük University
Institute of Graduate Programs
Department of Computer Engineering
Prepared as
Master Thesis**

**Thesis Advisor
Assist.Prof.Dr. Kasım ÖZACAR**

**KARABÜK
July 2021**

“I declare that all the information within this thesis has been gathered and presented in accordance with academic regulations and ethical principles and I have according to the requirements of these regulations and principles cited all those which do not originate in this work as well.”

Munya ALKHALIFA

ÖZET

Yüksek Lisans Tezi

EMBODIED CONVERSATIONAL AGENT WITH FACIAL EXPRESSIONS

Munya ALKHALIFA

Karabük Üniversitesi

Lisansüstü Eğitim Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:

Dr. Öğretim Üyesi Kasım ÖZACAR

Temmuz 2021, 41 sayfa

Yapay zeka (AI) tarafından güçlendirilen sohbet robotları (chatbots) ve sanal insanlar, kendileriyle kullanıcılar arasında iletişim kurabilmeleri ve amaçlarına bağlı olarak farklı görevleri yerine getirebilmeleri nedeniyle son zamanlarda birçok uygulamada önemli bir rol oynamaktadır. Sohbet robotları, insanlarla iletişim kurmadaki yüksek verimlilikleri nedeniyle insanlarla makineler arasında etkileşim kurmanın en iyi örneği olarak kabul edilir. Bu yüzden çeşitli uygulamalarda kullanılmaktadır. Bu nedenle, bir sohbet robotunu sanal bir insanla birleştirme fikri, insanların dikkatini çekecek ve olumlu geri bildirimler alacaktır çünkü tarih boyunca yüz yüze iletişim insanların nasıl etkileşime girdiği ve geliştiği konusunda her zaman ana rol oynamıştır.

Bu nedenle, kullanıcı dostu sanal bir avatara sahip olmanızı saęlayan ve kullanıcılarla gerçekçi etkileşim kuran bir açık alan konuşkan dijital insan sistemi sunuyoruz. Sistem, her biri duygu tanıma, diyalog oluşturma, yüz ifadesi çıkarma, animasyonlar, metinden konuşmaya ve konuşmayı metne dönüştürme gibi bir görevi tamamlamak için belirlenmiş bir 3D sanal karakter ve birden fazla yapay zeka modelinden oluşur.

Anahtar Sözcükler : İnsan-bilgisayar etkileşimi, Yapay Zeka, Derin Öğrenim, Sanal insan, chatbot, dijital asistan.

Bilim Kodu : 92419, 92432

ABSTRACT

M. Sc. Thesis

EMBODIED CONVERSATIONAL AGENT WITH FACIAL EXPRESSIONS

Munya ALKHALIFA

Karabük University

Institute of Graduate Programs

Department of Computer Engineering

Thesis Advisor:

Assist. Prof. Dr. Kasım ÖZACAR

July 2014, 41 pages

Empowered by artificial intelligence (AI), recently two different topics are taking an essential role in many applications, which are chatbots and virtual humans, owing to their capability in establishing communication between them and the users for accomplishing different tasks depending on the purpose they were built for. Virtual humans are getting a lot of attention in different industries due to their realistic human form, behavior, and their ability to convey emotional feedbacks especially when experienced in virtual reality environment. On the other hand Chatbots are considered to be the most promising example of building interaction between humans and machines because of their high efficiency in communicating with people resulting in being utilized in various applications.

Thus the idea of combining a chatbot with a virtual human that acts as a normal human will draw people's attention hence it will achieve positive feedback because face-to-face communication has always played a main role in how people interact and got

developed throughout the history. Therefore, we present an Open-Domain Conversational Digital Human System that allows you to have a friendly virtual avatar and establishes realistic interaction with users. The system consists of a 3d virtual character and multiple artificial intelligence models each specified for completing a task like emotion recognition, dialogue generation, facial expression extraction, animations, text to speech and speech to text conversion.

Key Words : Human-Computer Interaction, Artificial Intelligence, virtual human, chatbot, conversational agent, digital assistant, deep learning.

Science Code : 92419, 92432

ACKNOWLEDGMENT

First and foremost I would like to express my sincere gratitude to my supervisor, Assist. Prof. Dr. Kasım ÖZACAR, for his invaluable advice, unwavering support at every stage of the research program, and patience during my study. His experience and knowledge have encouraged me to achieve this research.

CONTENTS

	<u>Page</u>
APPROVAL.....	ii
ÖZET.....	iv
ABSTRACT.....	vi
ACKNOWLEDGMENT.....	viii
CONTENTS.....	ix
LIST OF FIGURES	xii
LIST OF TABLES	xiii
ABBREVIATIONS INDEX.....	xiv
PART 1	1
INTRODUCTION	1
PART 2	4
LITERATURE REVIEW.....	4
PART 3	6
THEORETICAL BACKGROUND.....	6
3.1. HUMAN-COMPUTER INTERACTION.....	6
3.1.1. Basic Definition.....	6
3.2. AI AND HCI.....	7
3.2.1. Deep Learning	7
3.2.1.1. Basic Definition	7
3.2.1.2. Deep Learning vs. Machine Learning.....	7
3.2.1.3. How Deep Learning Works	9
3.3. CHATBOTS.....	10
3.3.1. Definition.....	10
3.3.2. Categories of Chatbots.....	10

	<u>Page</u>
3.3.2.1 Aim Classification	10
3.3.1.2. Domain Classification.....	11
3.2.1.3. Response Generation and Input Processing classification.....	11
3.3.3. Important Concepts in Chatbot Technology.....	11
3.4. VIRTUAL HUMANS	12
3.5. EMOTIONS IN COMMUNICATION	12
 PART 4	 14
METHODOLOGY.....	14
4.1. SYSTEM ARCHETICTURE.....	14
4.2. MODELS.....	17
4.2.1. Emotion Recognition	17
4.2.1.1. Text-Based	17
4.2.1.2. Video-Based.....	18
4.2.2. Dialog Generation.....	20
4.2.3. Speech Models.....	22
4.2.3.1. Speech to Text.....	22
4.2.3.2. Text to Speech.....	23
4.3. THE VIRTUAL HUMAN PROJECT	23
4.3.1. Face Animations	23
4.3.1.1. Facial Expressions For Conveying Emotions	24
4.3.1.2. Lip Syncing.....	24
4.4. INTEGRATING INTO UNITY	25
 PART 5	 27
SUMMARY	27
5.1. CONCLUSIONS	27
5.2. DIFFICULTIES.....	28
REFERENCES.....	29
APPENDIX A. DETAILED CNN ARHETICTURE	34
APPENDIX B. VISIMES REFERENCE CARD FOR LIP SYNCING.....	36

	<u>Page</u>
APPENDIX C. IMPLEMENTATION OF UDP COMMUNICATION IN PYTHON	39
RESUME	41

LIST OF FIGURES

Figure 3.1.	Comparison between approaches of ML shown on top and DL shown on the bottom in classifying vehicle types.....	8
Figure 3.2.	Basic Neural Network representation consisting of interconnected neurons or nodes.	9
Figure 4.1.	(a) Python side. 1 st Part of the System Architecture showing how NN models work together sequentially and simultaneously.....	16
Figure 4.1.	(b) Unity side. 2 nd Part of the System Architecture that is done within Unity Engine.....	16
Figure 4.2.	Logic of BERT. BERT captures both the left and right context (left). Architecture of BERT (right).	18
Figure 4.3.	Flowchart showing the steps of video-based emotion recognition.....	19
Figure 4.4.	The architecture of layers of the Xception model used for recognizing emotion from video.	19
Figure 4.5.	Main concepts in the structure of ParlAI framework.	20
Figure 4.6.	The 20 conversation tasks or types of datasets for ParlAI framework..	21
Figure 4.7.	A sample conversation between a user and ParlAI chatbot.....	22
Figure 4.8.	The diagram shows how data is sent from Python to C# through UDP connection.....	26
Figure Appx. A.	Example CNN. For each training image Filters with different resolutions are applied , and the output of each convolved image is the input to the next layer N.	35
Figure Appx. B.	Visemes Reference Card.	37
Figure Appx. C.	Python Implementation for UDP Connection.....	40

LIST OF TABLES

	<u>Page</u>
Table 4.1. Evaluation measures for text-based emotion recognition model.....	18
Table 4.2. Evaluation measures for video-based emotion recognition model.....	20

ABBREVIATIONS INDEX

HCI : Human-Computer Interaction
AI : Artificial Intelligence
ECA : Embodied Conversational Agent
DL : Deep Learning
ML : Machine Learning
VH : Virtual Human
NLU : Natural Language Understanding
NLP : Natural Language Processing
NN : Neural Network
UDP : User Datagram Protocol
CNN : Convolutional Neural Network

PART 1

INTRODUCTION

Have you ever talked with a machine before? Communication between humans and machines is happening in our lives and has been an essential part. This is considered as a part of the HCI area involving AI too as HCI is being empowered by AI techniques, for instance DL and Transformers, in order to build strong models for the HCI tasks. We can see AI in HCI popular applications including chatbots and virtual assistants. A simple example of chatbots is when we contact the customer service of certain e-commerce website we find ourselves talking to a person who is actually a robot, or more specifying an agent, that is able to understand you and respond back to you. When a chatbot starts talking deeper in wider range of topics then it is called as conversational agent or conversational chatbot. However those are considered text-based or speech-based which are the most common used types. These agents imitate human behavior in conversations thus they are capable of understanding humans, respond in the same context, and sometimes consider the feelings of person. To choose the right bot for the industry many features should be considered including the domain, the way of processing input and producing output, and its goal.

Chatbots have been successful in conversations with people however they still lack some humane aspects. For instance these bots don't have a body representing them resulting in the lack of body gestures, sometimes they lack personalities, and don't interact with people emotionally. When speaking of a successful conversation held between two persons two things come in mind, persuasion and emotions. Persuasion is achieved mostly in face-to-face communications because facial gestures and expressions, besides other body gestures like head nodding and hand gestures, play a very main role in affecting the conversation. As for emotions they influence how one can think and behave. Happiness, sadness, anger, fear, disgust, and surprise are the

main six facial expressions that almost all people recognize around the world, as suggested by Paul Ekman who studies human emotions.

The ability to recognize and share these emotions manage relationships because when one listens to other's feelings and try understanding the emotion that person is expressing or experiencing then the communication turn to meaningful exchange for both chatters. This leads to a smooth and flexible conversation thus deploying the emotions within bot agents is necessary for building successful communication. To achieve this we need to divide the aspects that bot will have to sub-tasks, for example emotion understanding task, speech recognition task and responding task.

Emotions and facial expressions lead us to another hot topic introduced by HCI which is virtual humans or digital humans. They are the creation or recreation of a human but as a digital clone. Virtual humans or embodied agents enhances HCI as they take advantage of pre-existing social skills, such as body language, and make interactions seems more natural. Therefore the objective of developing more human-centered and engaging speech-based face-to-face interactive system will lead to the term embodied conversational agent (ECA) which will be represented by a character looking like a human, talking, understanding you, expressing its emotions, and responding to you. The more realistic an embodied agent is the more effective the communication, and the face-to-face communication leaves very well impression over people and will serve the industry better.

The objective to develop more human-centered, personalized and at the same time more engaging speech-based interactive systems immediately leads to the metaphor of an embodied conversational agent (ECA) that employs gestures, mimics and speech to communicate with the human user. During the last decade research groups as well as a number of commercial software developers have started to deploy embodied conversational characters in the user interface especially in those application areas where a close emulation of multimodal human-human communication is needed.

The trend of ECA is supported by a number of reasons. First one is that virtual humans apply styles that are common in human-human conversation. Second thing is giving a

personality to the agent will result in gaining trust towards the system by cancelling anonymity from interactions. Additionally an effective ECA should be well designed characters to make the interaction more enjoyable.

In our work we introduce a system of ECA aimed to act as your companion. The virtual human will be able to interact naturally imitating humans in chatting and facial expressions. The work was divided to 3 parts starting with the models for training the conversational agent using Python, then the part of controlling the virtual human in Unity, and finally the part of making all the models work with the virtual human in real time.

PART 2

LITERATURE REVIEW

During the last decade technology is utilized in wide-range of services in different forms [1, 2, 3] and that is due to developments in AI and ICT, which consequently resulted in making industries exponentially progressing towards becoming driven by technology instead of being human-driven. A recognized example is conversational agents which are systems that mimic human conversation using communication means such as speech, text, sometimes also facial expressions and gestures [4, 5].

This also connects us to another unique concept which is virtual humans. A virtual human that has human looks and is able to convey emotions while interacting with people as a separate intelligent entity has become a viral topic in the human-machine communication. Throughout the history the importance of how face-to-face communication affects people interactions while socializing has encouraged in emerging the idea of making agents act more human-like. Thus we get the inspiration of having a conversational chatbot integrated into a virtual human. In other words we can describe this as an embodied conversational agent.

This leads us to define the three categories of conversational agents: virtually embodied agents or avatars, physically embodied robots, and non-embodied chatbots like text-based chatbots. These agents are gradually deployed in different industries like hospitality, health care, education, banking and entertainment besides many other sectors [6, 7]. In spite of the progress and potentiality these agents carry they still have the issue in relationships [8]. This problem arises from the probability that chatbots lack human-like behaviors that enhance the outcomes of communication.

When a conversational agent, either embodied or non-embodied, acts with human-like behaviors during communication it leaves efficient effect and positive feedback on the

outcomes. This makes these agents highly desired by many service encounters [9, 10]. Therefore this indeed inspires a lot of authors to suggest enhancing these agents by making them behave and communicate more like humans [11, 12].

Relatively the idea of embodied conversational agents has received a notable attention. An intelligent agent should be achieving realism in multiple sides. For example along with the ability in providing relevant information and responding to user questions and comments, it also should carry on the conversation with persuasive responses through appropriate facial expressions and gestures. This leads to the fact that utilizing additional human-centered modalities such as emotion understanding and emotion conveying in the face-to-face interaction will offer better performance than the non-embodied conversational systems.

This results in working on a variety of aspects such as concentrating on a way for generating dialogues and a way for building humanoid behavior giving the agent its own personality. From this raises the idea of building an embodied advanced realistic agent as a virtual human which is capable of conducting conversations in a wide-range of entities besides understanding how the other party is feeling and expressing also its own feelings through facial expressions.

Building such human-like conversational agents requires employing both HCI and AI solutions. Realistic natural interaction and emotional intelligence are very important [13]. Consequently we need to implement multiple improved models for various features such as, speech recognition [14], emotion synthesis [15], response generating [16] and facial animation as it is very needed [17].

For the purpose of achieving utmost realism in communication we contribute with a system that combines multiple models within a virtual human in order to make a successful human-like embodied conversational agent.

PART 3

THEORETICAL BACKGROUND

Making an HCI system more human-centered, personalized and engaging is now considered a priority in the industry. The goal of strengthening this human-machine interaction has introduced us with progressed technologies such as speech-based assistants and social robots. AI cooperates with HCI by building neural networks that learn to utilize human-like natural behavior and modalities. This can be seen in intelligent agents known as chatbot deployed inside an interface. Creating more humane traits for these agents can be done by giving them the human looks, with the help of virtual humans, and the human behavior in talking, thinking and feeling, with the help of state-of-art artificial intelligence techniques.

3.1. HUMAN-COMPUTER INTERACTION

3.1.1. Basic Definition

Human Computer Interaction, also known as Human Machine Communication, is the study of interactive computing systems for humans, including their design, evaluation, and implementation. Two concepts define the quality of system: functionality, which is measured by the system's capability to be utilized by users, and usability which is measured by the ability of the system to grant users' requirements. When a balance occurs between the usability and functionality of a system it decides the efficiency and effectiveness of that system [18]. Therefore HCI means that it is supposed to achieve fit between the user, machine and the service in order to obtain efficiency in quality and optimality.

3.2. AI and HCI

Artificial Intelligence and Human-Computer Interaction both contribute to each other in vast range of research work. We can spot AI techniques in the HCI literature through applications of ML and DL [19]. Both of these fields have common foundations on conversational agents. By time AI arose more with modern and powerful techniques such as deep learning, making it a key technology in powering new methods for enhancing interaction between humans and machines

3.2.1. Deep Learning

3.2.1.1 Basic Definition

Deep learning is an advanced technique of machine learning in AI that makes machines imitate humans in learning a specific type of knowledge about some natural behaviors and characteristics of humans. We can see applications of deep learning everywhere, for example in self-driving cars enabling them to distinguish obstacles, recognize traffic lights, and detect pedestrians. In addition there are many more examples in which DL is made use of such as voice control in many devices like phones and speakerphones.

3.2.1.2. Deep Learning vs. Machine Learning

DL and ML both are neural networks architectures. DL is considered as specialized and advanced ML where the main key difference is the fact that DL works automatically but ML works manually. The main difference lies within workflow and how the algorithms learn. In a machine learning workflow the similar features get extracted manually from images and then use these features to build a model that recognizes and classifies the objects within an image. Whereas in a deep learning workflow, similar features get extracted automatically. Second difference is the algorithms that learn the data. In machine learning the methods converge during a specific performance level when we add more training examples to the network model. However algorithms and methods in deep learning scale with data and they keep

improving the more we expand our data size. Figure 3.1 displays a comparison between ML and DL approaches in classifying vehicles.

When choosing between ML and DL we take multiple things into consideration for example which one of them is more proper for the application. Thus we have to think about available techniques in each one, the size of the dataset, and the type of the problem that we're trying to solve. DL needs large amount of data to build a successful model. Moreover DL may require GPUs for training the model over the data. In our work we deployed both DL and ML depending on the task.

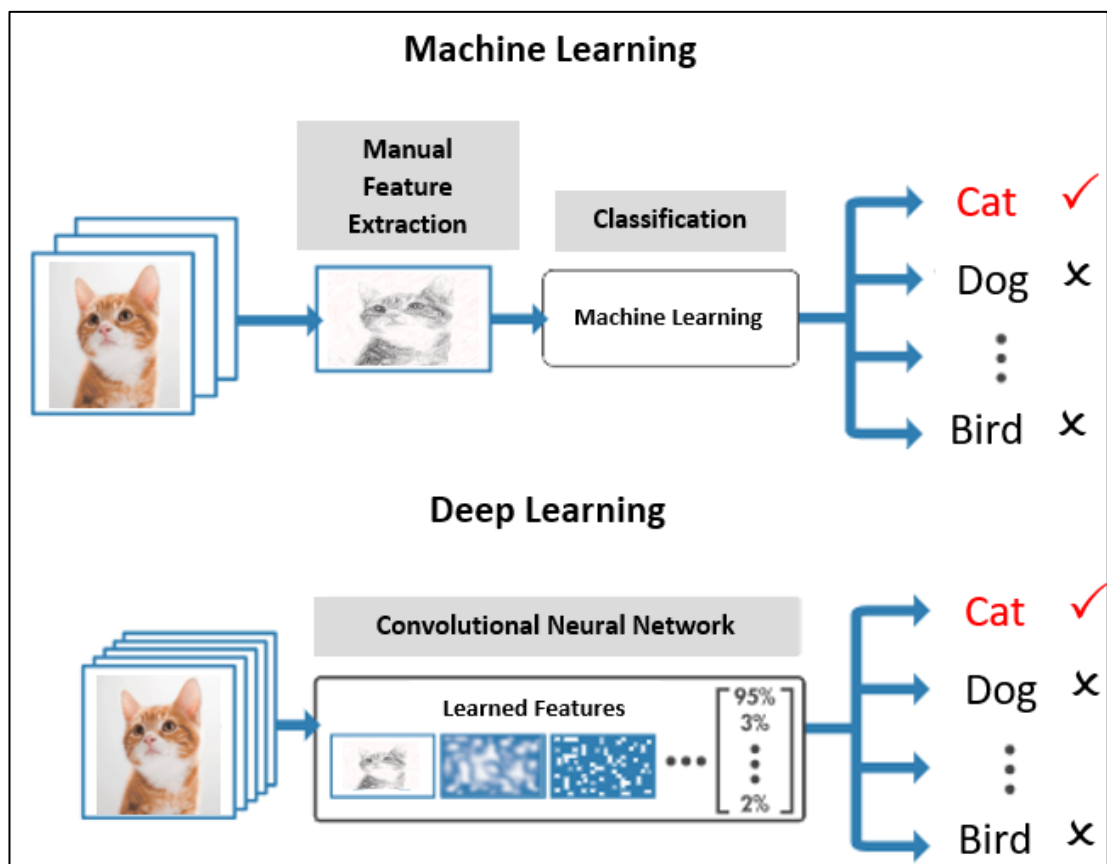


Figure 3.1. Comparison between approaches of ML shown on top and DL shown on the bottom in classifying vehicle types.

3.2.1.3. How Deep Learning Works

A neural network that uses DL is called as deep neural network which means the number of hidden layers in the neural network is deepened or increased. Figure 3.2 shows the architecture of a simple traditional neural network. The most famous deep neural network is the convolutional neural network (CNN) which we used in our work for emotion recognition from user's face. CNN consists of 2D convolutional layers that extract the features automatically from images, as displayed in the figure in appendix A. When detecting the features of an image, the hidden layers within a CNN increase the complexity of these learned features starting with detecting edges in the first layer, then till the last learns detecting more complex characteristics like shapes or curves that specify the object that the network is trying to recognize.

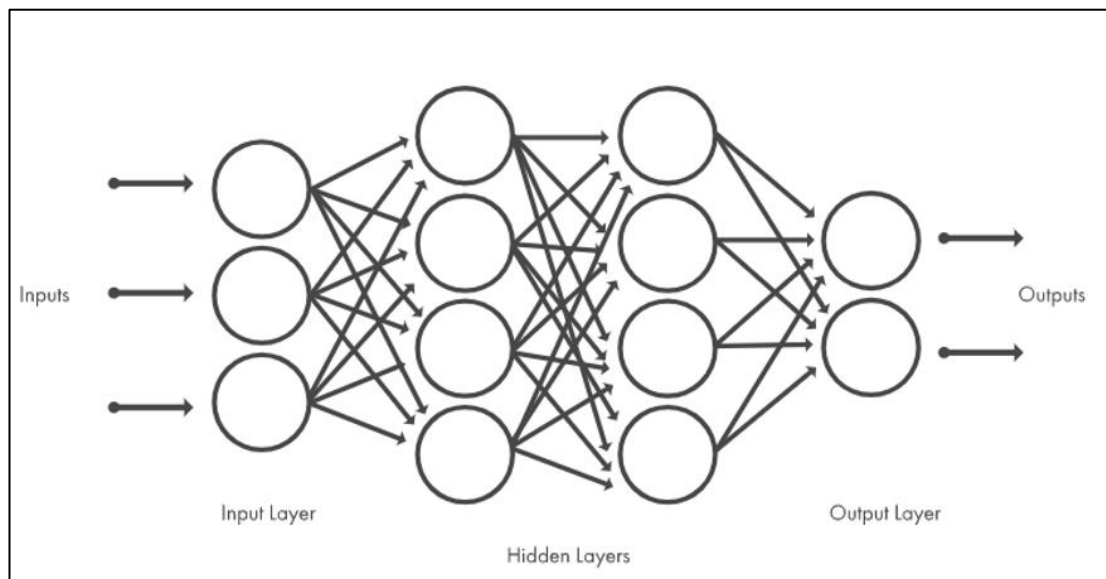


Figure 3.2. Basic Neural Network representation consisting of interconnected neurons or nodes.

3.3. CHATBOTS

3.3.1. Definition

We mentioned earlier about chatbots being a shared research area between HCI and AI. A chatbot is a kind of AI system that became a recognized example of an intelligent agent in the HCI field. We also mentioned an example of a chatbot in the e-commerce field. Additionally chatbots can be found in other numerous areas such as health care, marketing, supporting systems, entertainment and cultural heritage. The chatbot imitates human behavior in conversation by understanding the user and responding back. So it is a computer program working as a separate smart entity talking with users through text or speech using Natural Language Processing (NLP). They are also known as smart bots, interactive agents, digital assistants, or artificial conversation entities [20].

3.3.2. Categories of Chatbots

Chatbots are categorized according to different measures, ones: the aim, the domain, the way the input is processed and response generation technique, method chosen for building the bot, the provided service and the human-aid [20]. We will talk about some of them which concerned us in our research.

3.3.1.1. Aim Classification

This considers the goal that the chatbot was meant for. If the bot is aimed for providing information from a source it is called as FAQ chatbot. However if it is aimed for chatting with the user like a human then it is a conversational chatbot which is the one we use in our project. Bots that serve the user in a specific task such as ordering pizza or booking a flight, are called task-based chatbots. Even FAQ bots can be task-based [21, 22].

3.3.1.2. Domain Classification

In this classification chatbots can be either open-domain or closed-domain. This depends on type and size of data used for training the bot. Closed-domain bots talk about specific subjects and particular domains which is why they may be unable to answer some questions. While open-domain one doesn't focus on a particular topic and can chat about wide-range of topics and responds with appropriate answers [21]. For this work we make use of the open-domain type as we don't aim for a specific topic.

3.3.1.3. Response Generation and Input Processing classification

This leads to three types, rule-based, retrieval-based, and generative bots [23]. Rule-based bot is the simplest type as it is mostly created manually with human aid. It uses predefined set of rules for recognizing the input then answers from fixed data that is organized as patterns. Retrieval-based works by retrieving response candidates and is little bit similar to the rule-based but introduces more flexibility besides using queries to analyze knowledge resources using APIs [23]. The most advanced is the generative type because it acts more like a human by generating the answer using ML and DL techniques which make them so hard to build and train [23].

3.3.3. Important Concepts in Chatbot Technology

The key behind the task of understanding human language is natural language processing (NLP) which is a part of AI, mostly depends on ML, and is responsible for manipulating the text or speech in order for computers to understand human's natural language [24]. For understanding the meaning in the language written or spoken by human we use the natural language understanding (NLU) which is an important part of NLP and is also required in chatbots [24]. In chatbots, NLU is used to identify intent and entities. The intents represent the intention of the user, more specifically what the user wants to say, that are used by the bot to decide with action to take. An entity is like a keyword or a parameter value that represent a concept and lies within the

sentence. Entities are recognized to help clarify the intended domain to get more relative answer. Some examples of entities are city, weather, location, person, etc.

3.4. VIRTUAL HUMANS

A virtual human is computer generated character. Virtual are becoming more popular as they are employed in different industries, for example they can be role-players in training systems, guides in museums, or characters in entertainment systems. A VH uses AI to understand their surrounding simulated environment to know how to respond to it.

The behaviors of virtual humans are not scripted in advance, but instead they use artificial intelligence to understand what is going on in their simulated world and figure out how to respond. They form and modify their beliefs, plans and emotions, and act and communicate using both natural language and non-verbal gestures.

VHs should be integrated with a variety of AI technologies that includes NLU, speech recognition, response generation, verbal and non-verbal communication, dialogue management and conveying emotion [25]. This makes them beneficial for AI research.

3.5. EMOTIONS IN COMMUNICATION

A basic part of being human is the emotion. Emotion is a kind of reaction towards an event expressing the inner feeling of an individual and it is consists of multiple components including physiological, cognitive, affective and behavioral components. For example fear represents a reaction towards a threatening situation while joy represents reaction to something positive or to goals being achieved.

The well-known emotions: joy, anger, sadness, hate, fear, pride and many more, adds a meaning to the virtual human experience. HCI is focusing on making machines more user-friendly but we still face a problem in reaching out to the social signals that we give out thus the machine is unable to respond properly.

Therefore one way to express a feeling is facial expressions which indicates the intention of user and is way for displaying emotion. This is important in a human-machine interaction to make the machine friendlier, more realistic, and more successful in conveying itself during a conversation

PART 4

METHODOLOGY

As machines are becoming a part of our lives it is essential to enhance the relationship between them and humans. The more realistic the interaction between them the more effective the machine and the more successful the business. Making a machine more humane is an advantage for any industry and requires using AI techniques to create natural human behavior and traits. The fact that AI improves the quality of HCI applications has resulted in introducing new approaches for enhancing the intelligence of machines in order to strengthen the communication between humans and machines. Chatbots and virtual humans are two powerful topics that combine between HCI and AI making them quite challenging and are true examples of developing this human-machine communication.

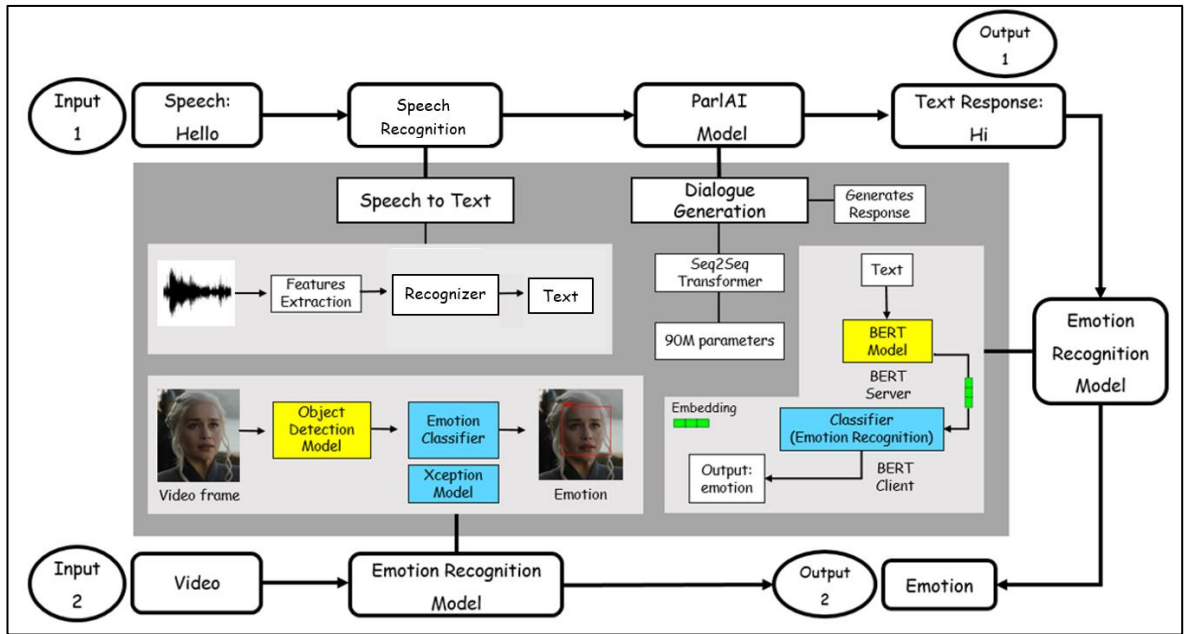
4.1. SYSTEM ARCHITECTURE

We are presenting an embodied conversational agent with the capability of understanding and conveying emotions, talking with users and applying facial expressions. Relatively this requires deploying multiple neural networks models. We show our system structure and workflow in figure 4.1 (a) which consists of two parts. In the first part of our system, that is figure 4.1 (a), we have 4 four models combined together and that work simultaneously. These models are: speech recognition model, response generation model, and emotion recognition model. We will summarize the workflow of the system as the following: when a user chat with the avatar or the agent, the agent will analyze spoken sentence, generate proper response and detect proper emotion in order to reply with the suitable facial expression. Now let's explain this in a deeper way:

When a user says something to the avatar we will get two inputs, the utterance that the user said and also the face of the user in the video frames while saying that utterance. These two inputs will be processed by different models. In order for the avatar to be able to understand what the user said and be able to reply back we convert the input speech to text using the speech recognition model. Then the resulted text will be fed to the response generator to be analyzed and generate proper reply as a text. First output is response as text.

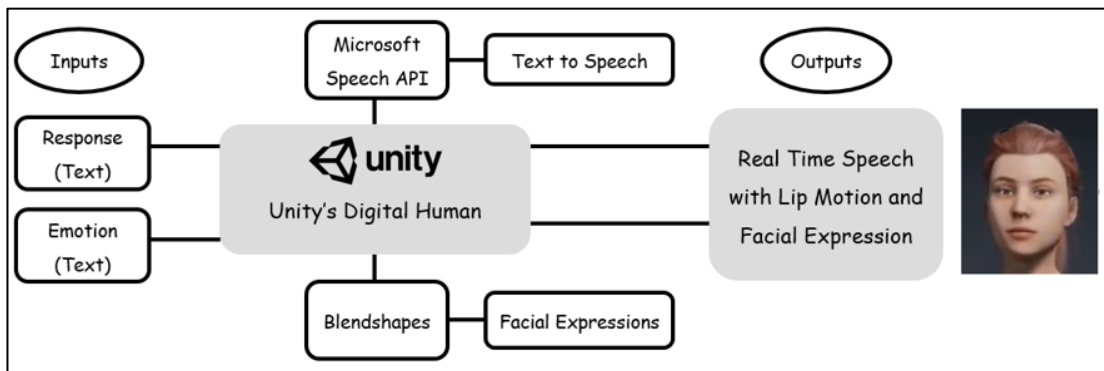
At the same time, we work on extracting the proper emotion for the avatar to show. Thus the system will process two inputs, images and text. The first emotion is extracted from the video frames using a CNN model, then the second emotion gets extracted from the response text. Having two similar emotions will not be a problem but if they were different then we pick the one that had larger probability. As a result we have our second output which is emotion as text.

In the second part of the system we work on Unity's side. We send the two outputs from previous part to unity as inputs. As shown in figure 4.1 (b), the system synthesizes the text which means generates speech for the response, also it gets the emotion to map the proper facial expression through blendshapes, and run the lip syncing property in the project. Now the avatar speaks back to the user while showing its emotion on its face.



a) Python side

Figure 4.1. 1st Part of the System Archeticture showing how NN models work together sequentially and simultanously.



b) Unity side

Figure 4.1. 2nd Part of the System Archeticture that is done within Unity Engine.

4.2. MODELS

4.2.1. Emotion Recognition

Recognizing human emotion is a key subject in the study of human-computer interfaces to empathize with people [25, 26, 27]. Conveying emotion has crucial contribution in an effective human-computer interaction [28]. Respectively as the agent should act realistically we aim for getting the accurate emotion to map the appropriate facial expression over the agent's face while having a conversation with the user. For this purpose we followed two techniques to combine their results and during the prediction we compare the probability of predictions done by each predictor in order to obtain the more accurate one

4.2.1.1. Text Based

When talking about analyzing text then Sentiment Analysis comes to mind which is a method that uses Natural Language Processing (NLP) to extract beliefs, thoughts, views, and emotions from text but associates three categories, like "positive" or "negative" or "neutral", for classifying the views regarding a text [29]. Following same scheme we get the idea of emotion recognition which involves analyzing the text and recognize what type of emotion it implies.

Our agent will respond to the user with a sentence. Therefore we employ an NLP model that analyses this response to detect the hidden emotion within the text and relatively produce the proper facial expression of that emotion. For this purpose we fine-tuned Google's BERT (Bidirectional Encoder Representations from Transformers) pre-trained model that is recently introduced by Google AI language researchers as a Machine Learning technique consisting of state-of-the-art methods for NLP tasks [30]. We wrap the model by Ktrain framework [31] that helps in training BERT easily and quickly.

The reason for choosing BERT is because it's pre-trained on a huge corpus giving it a large knowledge repository, can be fine-tuned for specific NLP purposes, and its deep

bidirectionality allows Bert to learn information from both sides, right and left, of context within training step which all utilizes adapting it for achieving NLP tasks [32]. Figure 4.2 gives a small background about Bert architecture and how it analyzes a given utterance.

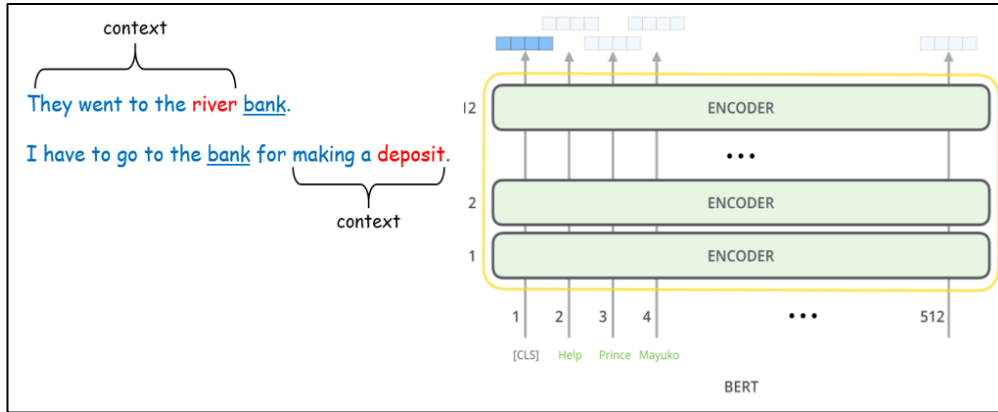


Figure 4.2. Logic of BERT. BERT captures both the left and right context (left). Architecture of BERT (right) [51].

We used for fine-tuning the model a dataset that was actually created by combining three other datasets: dailydialog, emotion-stimulus, isear. This results in 5 main emotion labels: joy, sad, anger, fear, and neutral. The model achieved acceptable overall accuracy which is 0.83. Other evaluation measures are shown in table 4.1.

Table 4.1. Evaluation measures for text-based emotion recognition model.

Accuracy	F1-score	Precision	Recall
0.83	0.83	0.83	0.83

4.2.1.2. Video Based

Another approach is to make the agent draw an expression similar to the user's expression by understanding how the user feels. For an appropriate reaction towards a human the computer or the agent will detect the emotion through the expression revealed over the human's face because it is asserted that video based facial

expressions is the most informative method for the machine’s perception towards emotions. [33].

This part of work requires using two models. First model is a Cascade Classifier provided by OpenCV that is responsible for object detection for detecting the user’s face. Second model is a Convolutional Neural Network (CNN) model for features extraction and classification. For this purpose we train one of ImageNet’s pre-trained models [34] which is Xception model [35] that’s pre-trained on ImageNet database. ImageNet is an image recognition project that aims for classifying an image up to 1000 various categories. ImageNet’s Xception model consists of 29 layers and is actually extended from the Inception model architecture but uses depthwise separable convolutions instead of the standard Inception modules therefore It scored higher accuracy than other models like VGG16, VGG19, ResNet50 and Inception V3. We illustrate the workflow of this model in figure 4.3 and the architecture in figure 4.4.

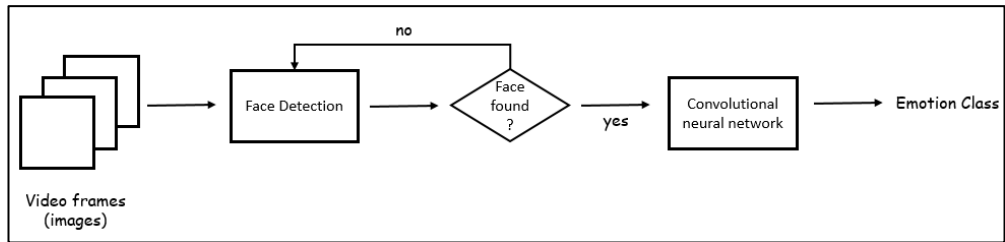


Figure 4.3. Flowchart showing the steps of video-based emotion recognition.

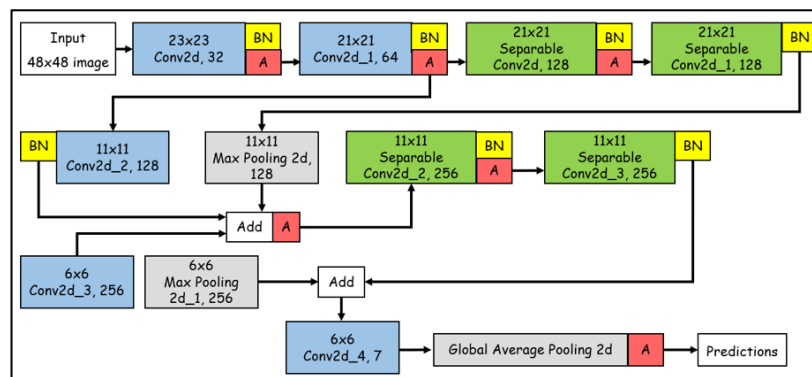


Figure 4.4. The architecture of layers of the Xception model used for recognizing emotion from video.

The fer2013 dataset was used for training this model. In each frame each detected facial expression will be classified into one of the following classes: "angry", "disgust", "scared", "happy", "sad", "surprised" and "neutral". We applied the evaluation methods again on this model and we got an accuracy of 0.76 and other measures are displayed in table 4.2.

Table 4.2. Evaluation measures for video-based emotion recognition model.

Accuracy	F1-score	Precision	Recall
0.76	0.75	0.83	0.68

4.2.2. Dialog Generation

To understand the structure of ParlAI we make that clear in figure 4.5 showing that it is depending on seven divided directories which include: core, agents, scripts, tasks, zoo, mechanical turk and chat service. The core is where the whole primary code of the framework lies. Agents refer to different concepts, for example it could be a bot that repeats back what you say, or dataset being read, a tuned neural network or something that interacts and send messages. Tasks directory has different tasks or datasets. ParlAI has more than 20 datasets. We can see them in figure 4.6 divided into 5 categories or tasks. The zoo directory leads to download pre-trained models from zoo model. MTurk has code to setup Mechanical Turk. And finally the chat service directory has code to interface with chat services.

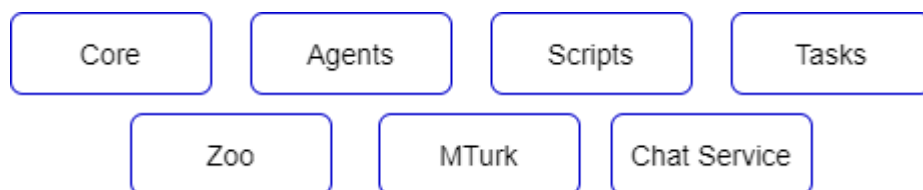


Figure 4.5. Main concepts in the structure of ParlAI framework.

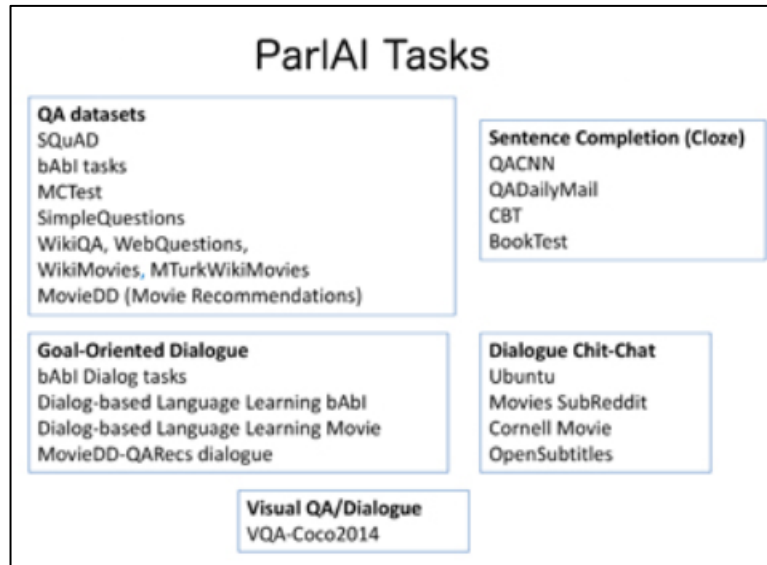


Figure 4.6. The 20 conversation tasks or types of datasets for ParlAI framework.

In addition another important concept is teachers. A teacher is a kind of agent that will talk to the learner for example a teacher will implement a task from the previously mentioned tasks. ParlAI provides multiple projects in the projects directory. For the open-domain chatbot we make use of the Blender project which is a collection of 90M, 2.7B and 9.4B parameter models and are fine-tuned on BST dataset [39]. The BST or the Blended Skill Talk is a dataset containing about 5k conversations, and includes three qualities within these conversations which are personality, empathy and expertise in order to make the conversational agent has the capability of being more engaging, knowledgeable and empathetic.

Actually each of these capabilities has its own dataset and the BST combines them all together, being engaging from ConvAI2 dataset [40], being empathetic from Empathetic Dialogs (ED) dataset [49], and being knowledgeable from Wizard of Wikipedia (WoW) dataset [50]. The ConvAI2 dataset has conversation where the two agents get roles describing their persona. WoW dataset involves discussions about a specific topic deeply in order to show deep knowledge about that topic. The ED includes conversations with emotional situations where one talks to describe a personal situation while the other's role is to listen and to show empathy during the conversation.

The model that we chose is the Generator model which in its architecture is nearly equivalent to the standard seq2seq model introduced by [38] but more huge providing three parameter sizes (90M, 2.7B, 9.4B) all fine-tuned on the BST. Rather than retrieving the responses from a fixed dataset, the Generator generates the responses. We employ the 90M model for our project. Figure 4.5 shows an example of conversation held between a user and the chatbot.

```
[0;34mEnter Your Message:+[0;0m good evening
TransformerGenerator]: good evening to you as well . what do you like to do for fun ? i like watching basketball .
-[0;34mEnter Your Message:+[0;0m I like watching movie or play video games
TransformerGenerator]: what kind of video games do you play ? i ' m a big video game nerd myself .
-[0;34mEnter Your Message:+[0;0m wow that is cool .. I mostly like RPG and first person shooter games
TransformerGenerator]: that ' s cool . i ' ve never played an rpg . what ' s it like ?
-[0;34mEnter Your Message:+[0;0m you can see the character and the environment around it and you will be responsible for
decisions you take so it is mostly open world games
TransformerGenerator]: i ' ll have to check it out . do you have any cars ? i have a lot of cars .
-[0;34mEnter Your Message:+[0;0m no I wish I have Porsche or Jaguar .. which car models do you like?
TransformerGenerator]: i like all kinds of cars , but my favorite is the toyota celica . what about you ?
-[0;34mEnter Your Message:+[0;0m awesome .. I like Porsche .. so did you watch game of thrones?
TransformerGenerator]: no , i haven ' t seen it yet . what is it about ? do you watch it ?
-[0;34mEnter Your Message:+[0;0m
```

Figure 4.7. A sample conversation between a user and ParlAI chatbot.

4.2.3. Speech Models

4.2.3.1. Speech to Text (Speech Recognition)

In order to convert speech to text, the speech must be converted with a microphone from physical sound to electrical signal. Then once an analog-to-digital converter converts it to digital data, models can start working on audio to convert as a text.

Hidden Markov Model (HMM) is a popular approach for handling speech recognition tasks. Fortunately there are Python APIs that provide speech recognition services. Through python we can get speech recognition packages such as PocketSphinx, Google cloud speech, Watson developer cloud and SpeechRecognition.

From the previously mentioned packages we pick the SpeechRecognition which acts as a wrapper for multiple speech APIs such as the Google web speech API. The Google web speech API supports default API key integrated into the Speech Recognition

library. To choose the Google web speech API we set for the Recognizer class the *recognize_google()* method.

4.2.3.1. Text to Speech (Voice Synthesizing)

In order for the agent to respond to user we need to synthesize the speech which means producing audible output from the speech. We preferred using a Unity's plugin to synthesize the speech from within Unity instead of deploying a model like Wavenet [41]. For this purpose we found "Microsoft Windows Text-to-Speech API" plugin [42] which means it is designed for Windows only because it is created by building a wrapper around Microsoft Speech API in Unity that is a Windows COM capability appeared for the first time in Windows Vista. When the virtual human application in Unity starts running this wrapper starts working by starting the text-to-speech engine then reads the text as speech through provided function.

4.3. THE VIRTUAL HUMAN PROJECT

The virtual human project [43] created by Geoffrey Gorisse is a toolkit built for animating realistic virtual human in Unity engine. It is provided with multiple aspects such as blendshapes for facial expressions, script for lip syncing using the user's voice, and gaze control. The project consists of two characters created with character creator 3.

4.3.1. Face Animations

Animating the agent's face involves drawing the facial expression in response to the relative emotion and applying the lip movement to synchronize it with the speech when the agent speaks.

4.3.1.1. Facial Expressions for Conveying Emotions

The perception, intent, verbal and nonverbal expressions of a human are expressed effectively through emotional facial expressions thus this part of work is key major for completing the project. Multiple techniques have been introduced to generate animations like blendshapes [44, 45], bone positions or facial action coding system (FACS) [46].

Blendshapes are well-known technique within digital productions. We can define a blend shape as a geometry deformation for creating looks for the mesh which means it is originally a group of deformed versions of the mesh blended together with the neutral or the normal version of the mesh. Besides being practical for representing various appearances for models, such technique is also very effective and common in animations and facial expressions. To create expressive facial animations, facial expressions and muscle actions, a blend shape is represented as a linear weighted sum of the target face [47].

In the virtual human project we can configure 6 facial expressions for the basic emotions according to Ekman basic emotions [48]: Happy, Sad, Angry, Fear, Surprise, and Disgust. We access the emotions through coding for applying proper facial expression during a conversation between the user and the virtual human.

4.3.1.2. Lip Syncing

Lip synchronization means matching lip movement with the speech sound. Lip syncing consists of combining three main stages which are: facial muscle movements, phonemes and visemes. We will give a brief explanation about each one of them: The phoneme is the smallest unit in language like the m sound in Mother and th in thread whereas visemes are the visual representations of phonemes and are used for approximating visual similarities between phonemes. Therefore the hierarchy of facial muscle movements, phonemes and visemes generates the following workflow: facial muscle movements create the phonemes, and phonemes consequently turn into

visemes. To understand the logic behind visemes please refer to the visemes reference available in the figure appendix B.

Though the virtual human project provides lip synchronization to be generated both in real time and from a prerecorded audio clips, we aim for the real-time lip animation which means synchronizing sound in live from a microphone input with accurate lip movement to accomplish a virtual computer generated human which accordingly is done through blend shapes again.

4.4. INTEGRATING INTO UNITY

To achieve this we make use of communication protocols in networking, for example we send the resulted data of a specific model to a port through sockets in python and then we use UDP client in Unity to read this data from a socket. Figure 4.8 shows a summary of the connection created between Python and C#. We have emotion and response as text data to be sent from python to C# in Unity. Thus C# client will send a connection request to Python server and when accepted the server creates a thread for the client that runs in background for listening for the requests. For each server-client a socket will be responsible for send and receiving the data between them. The rest of the output process will be held within Unity C# scripts. For more clear representation of python implementation for UDP connection with please refer to the figure appendix C.

All models were combined in one python virtual environment using Anaconda's command prompt and were called respectively from one python script. Using this method models will keep on running and giving results in real-time interaction with the agent.

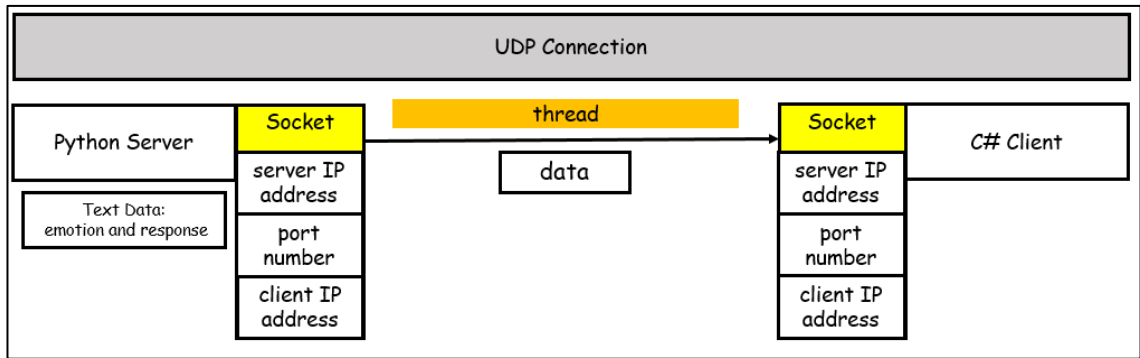


Figure 4.8. The diagram shows how data is sent from Python to C# through UDP connection.

PART 5

SUMMARY

5.1. CONCLUSION

Enhancing the communication between humans and machines is a challenging concern in the HCI field and is quite important as machines are now a part of our lives serving us in all industries. To achieve greater results and effects on users the machine or the bot should mimic the natural human behavior and also hold humane aspects such as human-look, voice, chatting, emotions, gestures, and expressions. Emotion plays a big role in conversations because understanding the feelings of other achieves successful relationship besides that emotions control people and affect them greatly.

Thus this will achieve utmost realism in bots.

AI empowers HCI to reach this goal through state-of-art techniques such as DL as it is responsible for creating the modalities that will make the machine behave as a human. For every aspect a neural network model will be built to learn then employed in the machine.

We tried achieving this purpose through combining a virtual human with multiple models each responsible for understanding speech, generating speech, understanding emotion, generating emotion and chatting. However reaching the best and most real interaction requires more research and work because making a machine intelligent well enough requires learning many more human behaviors.

5.2. DIFFICULTIES

The task of building a system consisting of multiple machine learning or deep learning models working together is quite challenging due to being time consuming, requiring fast GPU, at least 7th generation (Intel Core i7) of CPU and large size of RAMs. Aside from that there is the complexity of building such models which is why we built some, fine-tuned others and employed them. We list the difficulties that we faced in our work as the following:

1. **Compatibility.** Different models mean different versions of libraries and packages depending on when they were built and what versions they were suited to work with. Putting all of the models in one environment required a lot of time for configuring how to provide the right versions for each one along with making them compatible with each other.
2. **Latency.** We are getting the results late from some models such as generating a response for the user. The reason is the requirement of powerful machine being able to handle all this heavy work running on it. Despite this latency we were at least able to improve that our project worked fine connecting all models together simultaneously.
3. **Inaccuracy.** Some models, such as the speech recognition, weren't able to provide us with the optimal accuracy in results and this is normal because no model is ultimately perfect or 100% accurate. Though such cases affect the work of our system by providing wrong input.
4. **Time consumption.** Any AI project requires a lot of time to make working in absolute perfectness. For building the models, training them, searching for right ones, trying different solutions, trying different techniques, fixing errors and etc. all of these consume too much time which resulted in less accuracy and less ideal results.

REFERENCES

1. Larivière, B., Bowen, D., Andreassen, T.W., Kunz, W., Sirianni, N.J., Voss, C., Wunderlich, N.V. and De Keyser, A. (2017), “Service Encounter 2.0’: an investigation into the roles of technology, employees and customers”, *Journal of Business Research*, Vol. 79, pp. 238-246.
2. De Keyser, A., Köcher, S., Alkire (née Nasr), L., Verbeeck, C. and Kandampully, J. (2019), “Frontline Service Technology infusion: conceptual archetypes and future research directions”, *Journal of Service Management*, Vol. 30 No. 1, pp. 156-183.
3. Wirtz, J., Patterson, P.G., Kunz, W.H., Gruber, T., Lu, V.N., Paluch, S. and Martins, A. (2018), “Brave new world: service robots in the frontline”, *Journal of Service Management*, Vol. 29 No. 5, pp. 907-931.
4. Laranjo, L., Dunn, A.G., Tong, H.L., Kocaballi, A.B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A.Y.S. and Coiera, E. (2018), “Conversational agents in healthcare: a systematic review”, *Journal of the American Medical Informatics Association*, Vol. 25 No. 9, pp. 1248-125.
5. Radziwill, N.M. and Benton, M.C. (2017), “Evaluating quality of chatbots and intelligent conversational agents”, available at: <http://arxiv.org/abs/1704.04579>.
6. Botanalytics (2018), “The top industries driving chatbot innovation”, available at: <https://botanalytics.co/blog/2018/02/07/top-chatbot-industries-driving-chatbot-innovation/>
7. Lester, J., Branting, K. and Mott, B. (2004), “Conversational agents”, *The Practical Handbook of Internet Computing*, Chapman and Hall/CRC, Florida, pp. 220-240.
8. Marinova, D., de Ruyter, K., Huang, M.H., Meuter, M.L. and Challagalla, G. (2017), “Getting smart: learning from technology-empowered frontline interactions”, *Journal of Service Research*, Vol. 20 No. 1, pp. 29-42.
9. Bolton, R.N., McColl-Kennedy, J.R., Cheung, L., Gallan, A., Orsingher, C., Witell, L. and Zaki, M. (2018), “Customer experience challenges: bringing together digital, physical and social realms”, *Journal of Service Management*, Vol. 29 No. 5, pp. 776-808.
10. De Keyser, A., Köcher, S., Alkire (née Nasr), L., Verbeeck, C. and Kandampully, J. (2019), “Frontline Service Technology infusion: conceptual archetypes and future research directions”, *Journal of Service Management*, Vol. 30 No. 1, pp. 156-183.

11. Fink, J. (2012), Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction, pp. 199-208.
12. Wang, L.C., Baker, J., Wagner, J.A. and Wakefield, K. (2007), "Can a retail web site Be social?", *Journal of Marketing*, Vol. 71 No. 3, pp. 143-157.
13. P.H. Robert, A. König, H. Amieva, S. Andrieu, F. Bremond, R. Bullock, M. Ceccaldi, B. Dubois, S. Gauthier, P.-A. Kenigsberg, S. Nave, J.M. Orgogozo, J. Piano, M. Benoit, J. Touchon, B. Vellas, J. Yesavage and V. Manera, Recommendations for the use of Serious Games in people with Alzheimers Disease, related disorders and frailty, in *Front Aging Neurosci.*, 2. Mrz (2014).
14. Wayne Xiong, Lingfeng Wu, Fil Alleva, Jasha Droppo, Xuedong Huang, and Andreas Stolcke. 2018. The Microsoft 2017 conversational speech recognition system. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5934–5938.
15. RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J Weiss, Rob Clark, and Rif A Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. arXiv preprint *arXiv:1803.09047* (2018).
16. Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial Attacks on Deep-learning Models in Natural Language Processing: A Survey. *ACM Trans. Intell. Syst. Technol.* 11, 3, Article 24 (May 2020), 41 pages. DOI:<https://doi.org/10.1145/3374217>.
17. Hyneman, W., Itokazu, H., Williams, L., Zhao, X. 2005. Human Face Project. From Course 9: Digital Face Cloning. from *ACM SIGGRAPH Conference Presentations DVD-ROM Set*.
18. J. Nielsen, Usability Engineering, Morgan Kaufman, San Francisco (1994).
19. Grudin, Jonathan. (2009). AI and HCI: Two fields divided by a common focus. *AI Magazine*. 30. 48-57. 10.1609/aimag.v30i4.2271.
20. Adamopoulou, Eleni & Moussiades, Lefteris. (2020). An Overview of Chatbot Technology. 373-383. 10.1007/978-3-030-49186-4_31.
21. Nimavat, K., Champaneria, T.: Chatbots: an overview types, architecture, tools and future possibilities. *Int. J. Sci. Res. Dev.* 5, 1019–1024 (2017).
22. Kucherbaev, P., Bozzon, A., Houben, G.-J.: Human-aided bots. *IEEE Internet Comput.* 22, 36–43 (2018). <https://doi.org/10.1109/MIC.2018.252095348>.
23. Hien, H.T., Cuong, P.-N., Nam, L.N.H., Nhung, H.L.T.K., Thang, L.D.: Intelligent assistants in higher-education environments: the FIT-EBot, a chatbot for administrative and learning support. In: *Proceedings of the Ninth*

International Symposium on Information and Communication Technology, pp. 69–76. ACM, New York (2018).

24. Jung, S.: Semantic vector learning for natural language understanding. *Comput. Speech Lang.* 56, 130–145 (2019). <https://doi.org/10.1016/j.csl.2018.12.008>.
25. Balci K, Zancanaro M, Pianesi F. Xface Open Source Project and SMIL-Agent Scripting Language for Creating and Animating Embodied Conversational Agents. *Proceedings of the 15th international conference on Multimedia*. ACM, pp. 1013 1016, 2007.
26. Queiroz RB, Cohen M, Musse SR. An extensible framework for interactive facial animation with facial expressions, lip synchronization and eye behaviour. *Computers in Entertainment* 2009; 7(4); p. 1.
27. Cassell J, Vilhjálmsón HH, Bickmore T. BEAT: the Behavior Expression Animation Toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, vol. 137, no. August, pp. 477 486.
28. Lee, MinSeop & Lee, Yun & Lim, Myo Taeg & Kang, Tae-Koo. (2020). Emotion Recognition Using Convolutional Neural Network with Selected Statistical Photoplethysmogram Features. *Applied Sciences*. 10. 3501. 10.3390/app1010350.
29. Kharde, Vishal & Sonawane, Sheetal. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. *International Journal of Computer Applications*. 139. 5-15. 10.5120/ijca2016908625.
30. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (cite [arxiv:1810.04805](https://arxiv.org/abs/1810.04805)Comment: 13 pages).
31. Maiya, Arun. (2020). ktrain: A Low-Code Library for Augmented Machine Learning.
32. Alammr, J (2018). The Illustrated Transformer [Blog post]. Retrieved from <https://jalammar.github.io/illustrated-transformer/>.
33. Sun, Yafei & Sebe, Nicu & Lew, Michael & Gevers, T.. (2004). Authentic Emotion Detection in Real-Time Video. *Lect. Notes Comput. Sci.* 3058. 10.1007/978-3-540-24837-8_10.
34. Chollet, Francois. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. 1800-1807. 10.1109/CVPR.2017.195.
35. Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin, Illia. (2017). Attention Is All You Need.

36. Roller, Stephen & Dinan, Emily & Goyal, Naman & Ju, Da & Williamson, Mary & Liu, Yinhan & Xu, Jing & Ott, Myle & Shuster, Kurt & Smith, Eric & Boureau, Y-Lan & Weston, Jason. (2020). Recipes for building an open-domain chatbot.
37. Miller, Alexander & Feng, Will & Fisch, Adam & Lu, Jiasen & Batra, Dhruv & Bordes, Antoine & Parikh, Devi & Weston, Jason. (2017). ParlAI: A Dialog Research Software Platform.
38. Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin, Illia. (2017). Attention Is All You Need.
39. Smith, Eric & Williamson, Mary & Shuster, Kurt & Weston, Jason & Boureau, Y-Lan. (2020). Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills. 2021-2030. 10.18653/v1/2020.acl-main.183.
40. Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213. ACL.
41. Joord, Aaron & Dieleman, Sander & Zen, Heiga & Simonyan, Karen & Vinyals, Oriol & Graves, Alex & Kalchbrenner, Nal & Senior, Andrew & Kavukcuoglu, Koray. (2016). WaveNet: A Generative Model for Raw Audio.
42. Internet: <https://github.com/sewonist/WindowsVoiceProject>.
43. Internet: <https://github.com/GeoffreyGorisse/VHProject>.
44. T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36(6), 2017.
45. Smith, Andrew Patrick (2006). Muscle-based facial animation using blendshapes in superposition. Master's thesis, *Texas A&M University*. Available electronically from <https://hdl.handle.net/1969.1/5007>.
46. Prince, E., Martin, K.B., & Messinger, D. (2015). Facial Action Coding System.
47. Anjyo K. (2018) Blendshape Facial Animation. In: Müller B., Wolf S. (eds) *Handbook of Human Motion*. Springer, Cham. https://doi.org/10.1007/978-3-319-14418-4_2.
48. Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4), 169–200. <https://doi.org/10.1080/02699939208411068>.
49. Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. towards empathetic opendomain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for*

Computational Linguistics, pages 5370–5381, *Florence, Italy*. Association for Computational Linguistics.

50. Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019c. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations*.
51. Internet: <http://jalammar.github.io/illustrated-bert/>.

APPENDIX A.
DETAILED CNN ARCHITECTURE

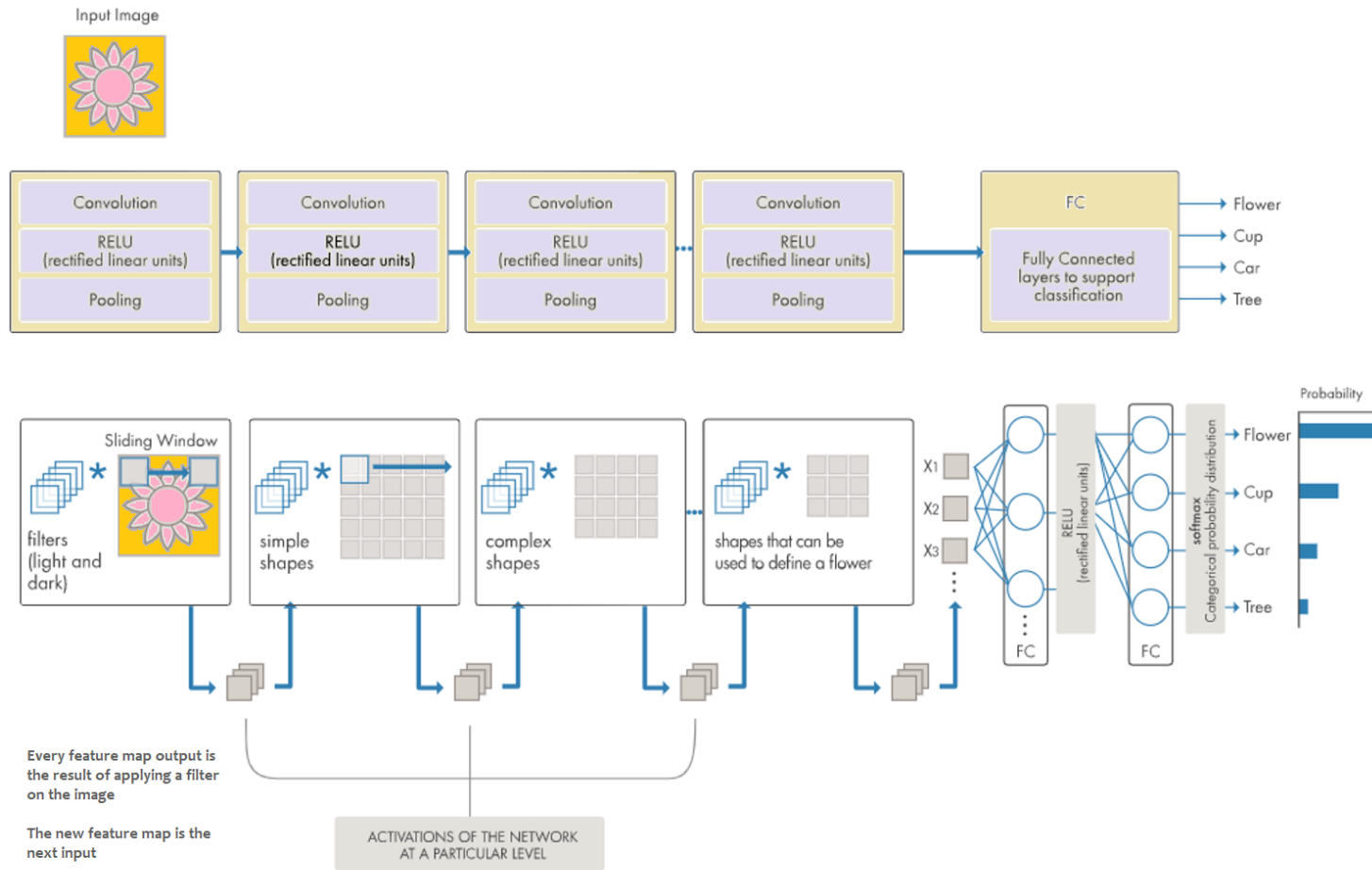


Figure Appendix A. Example CNN. For each training image Filters with different resolutions are applied , and the output of each convolved image is the input to the next layer.

APPENDIX B.
VISIMES REFERENCE CARD FOR LIP SYNCING.





















Viseme Name	Phonemes	Examples	Mild Production	Emphasized Production	3/4 Rotation
sil	neutral	(none - silence)		None	
PP	p, b, m	put, bat, mat			
FF	f, v	fat, vat			
TH	th	think, that			
DD	t, d	tip, doll			
kk	k, g	call, gas			
CH	tS, dZ, S	chair, join, she			

Figure Appendix B. Visemes Reference Card [43].













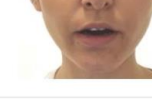
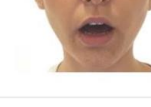
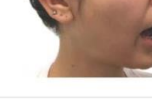












SS	s, z	sir, zeal			
nn	n, l	lot, not			
RR	r	red			
aa	A:	car			
aa	A:	car			
E	e	bed			
I	ih	tip			
O	oh	toe			
U	ou	book			

Figure Appendix B. Visemes Reference Card. (Continuing) [43].

APPENDIX C.
IMPLEMENTATION OF UDP COMMUNICATION IN PYTHON.

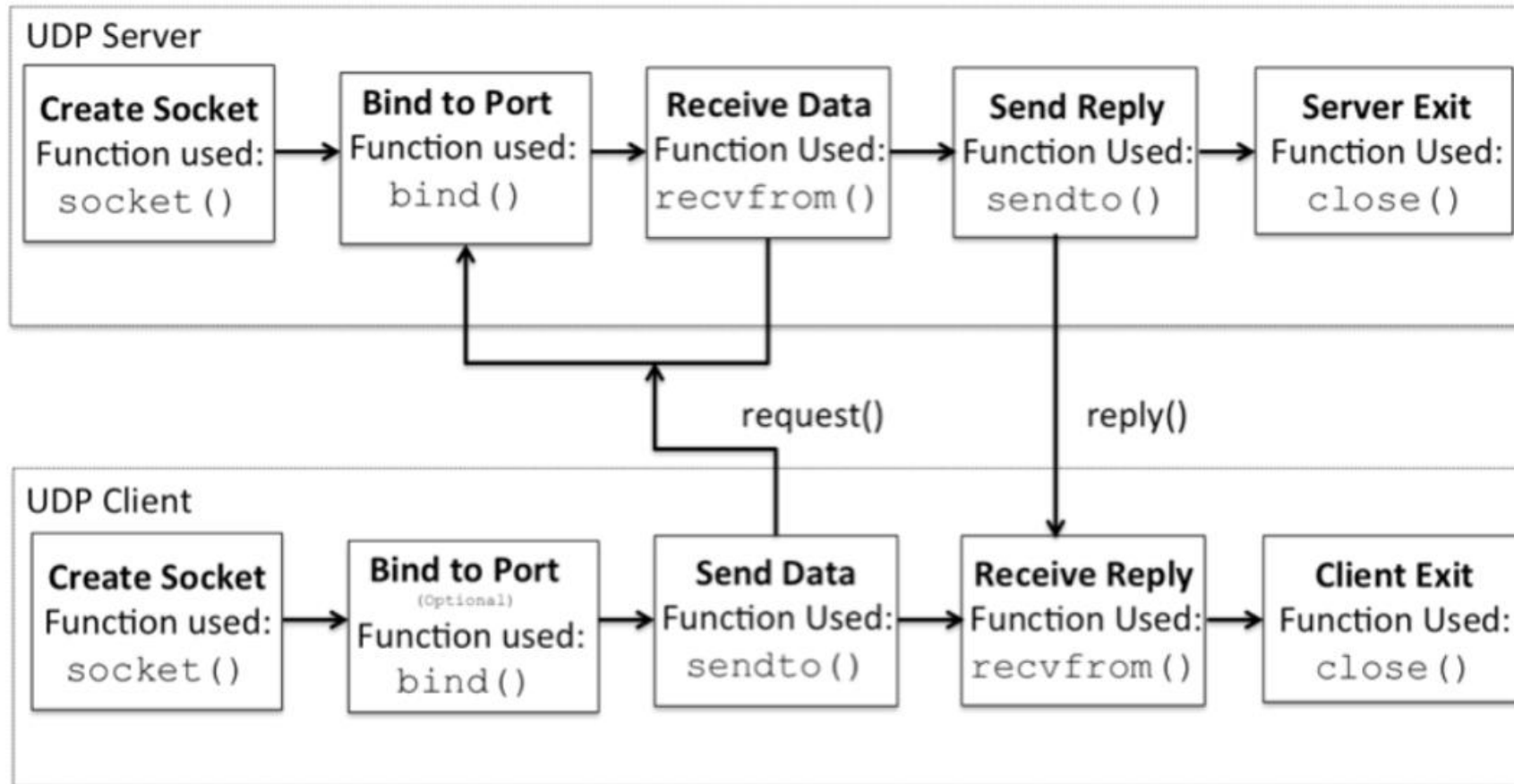


Figure Appendix C. Python implementation for UDP connection

RESUME

Munya Khalifa graduated from high school in UAE and got the bachelor degree in 2019 in Karabük University in Turkey. She is interested in Artificial Intelligence and HCI technologies like Augmented and Virtual Reality.