Turkish Journal of Science & Technology
19(1), 25-37, 2024

Research Paper
https://doi.org/10.55525/tjst.1301324

DigiHuman: A Conversational Digital Human with Facial Expressions

Kasım ÖZACAR[1*], Munya ALKHALIFA[2]
[1, 2] Computer Engineering Department, Engineering Faculty, Karabuk University, Karabuk, Turkiye
[*1] kasimozacar@karabuk.edu.tr, [2] munia.khalifa@gmail.com

(Geliş/Received: 24/05/2023; Kabul/Accepted: 06/09/2023)

Abstract: Recently, Artificial Intelligence (AI)-powered chatbots and virtual humans have assumed significant roles in various domains due to their ability to interact with users and perform tasks based on their intended purpose. Virtual humans have received considerable attention in various industries due to their lifelike human appearance, behaviour, and ability to convey emotions, especially in virtual reality contexts. Conversely, chatbots are finding use in a wide range of applications and represent a promising feature of human-computer interaction due to their efficient communication with humans. Therefore, this study aims to develop a real-time chatbot that can effectively convey emotions through facial expressions, thereby promoting realistic communication. To achieve this, several advanced AI models were employed to address different aspects, including speech recognition, emotion synthesis, and response generation. The methodology, models used, components, and results are explained in detail, and the results of the user study are also presented.

Keywords: human-computer interaction, artificial intelligence, virtual human, chatbot, conversational agent.

DigiHuman: Yüz İfadeleri ile Konuşan Dijital Bir İnsan

Öz: Yapay zeka destekli sohbet robotları ve sanal insanlar, oluşturulma amaçlarına bağlı olarak farklı görevleri yerine getirmek için kullanıcılarla aralarında iletişim kurma yetenekleri nedeniyle son zamanlarda birçok uygulamada önemli rol üstlenmişlerdir. Sanal insanlar, gerçekçi insan formları, davranışları ve özellikle sanal gerçeklik ortamında deneyimlendiklerinde duygusal geri bildirim iletme yetenekleri nedeniyle farklı sektörlerde büyük ilgi görmektedir. Diğer taraftan, sohbet robotları insanlarla iletişim kurmadaki yüksek verimlilikleri nedeniyle insan bilgisayar etkileşimi için en umut verici örneklerden biri olarak çeşitli uygulamalarda kullanılmaktadır. Bu nedenle, bu çalışmada başarılı bir iletişim ve gerçekçi davranış sergilemesi için yüz ifadeleri aracılığıyla duyguları iletme yeteneğine sahip bir gerçek zamanlı sohbet robotu oluşturması amaçlanmıştır. Bunun için sırasıyla; konuşma tanıma, duygu sentezi, yanıt üretme gibi çeşitli özellikler için çoklu geliştirilmiş yapay zeka modelleri uygulanmıştır. Çalışma kapsamında yaklaşım, kullanılan tüm modeller, bileşenler ve sonuçları kapsamlı bir şekilde açıklanmış ve kullanıcı testleri sonuçları da açıklanmıştır.

Anahtar kelimeler: insan-bilgisayar etkileşimi, yapay zeka, sanal insan, chatbot, konuşma ajanı.

1. Introduction

Metaverse technology is expected to contribute to and shape numerous substantial areas of our lives, including social media, commerce, education, and entertainment, and consequently, Metaverse requires intelligent conversational agents capable of delivering human-like responses to serve users in this evolving digital landscape. Furthermore, these agents are expected to imitate human behaviour within conversations, enabling them to comprehend human language, engage in responses within the appropriate context, and even grasp his/her feelings. Although numerous conversational agents or chatbots designed for conversation are available, they typically revolve around text or speech interactions and often lack certain human-like qualities. For instance, these bots lack a physical embodiment, leading to the absence of bodily gestures. Additionally, they may lack distinct personalities and fail to engage with individuals on an emotional level.

When considering an effective exchange between two people, persuasion and emotion come to mind. Persuasion is achieved primarily in face-to-face communications because the conversation mainly affects facial gestures and expressions, besides other body gestures like head nodding and hand gestures. Accordingly, an intelligent agent should be able to provide relevant information and respond to user questions and comments. It should conduct communication through persuasive responses with appropriate facial expressions and gestures. Emotions and facial expressions lead us to a hot topic: Human-Computer Interaction (HCI) introduced: virtual humans or digital humans. Virtual humans or embodied agents enhance HCI by taking advantage of pre-existing social skills, such as body language, and making interactions seem more natural. Therefore, the objective of developing a more human-centered and engaging speech-based face-to-face interactive system will lead to the term Embodied Conversational Agent (ECA). ECA will be represented by a character looking like a human, talking, understanding, expressing emotions, and responding to you. The more realistic an embodied agent is, the

[*] Corresponding author: kasimozacar@karabuk.edu.tr. ORCID Number of authors: [1] 0000-0001-7637-0620, [2] 0000-0003-0364-201X

more influential communication and face-to-face communication leave a very good impression on people and will serve the industry better.

Building such a high-fidelity agent requires employing both HCI and AI solutions. Realistic natural interaction and emotional intelligence are essential [1]. Consequently, we need to implement multiple improved models for various features such as speech recognition [2], emotion synthesis [3], response generating [4], and facial animation as they are very needed [5]. Despite creating or employing Convolutional Neural Network (CNN) and Natural Language Processing (NLP) models, in this paper, we concentrate on the problem of building human-agent real-time interaction with the ability to convey emotions through facial expressions to establish successful communication and realistic behaviour. The paper's contributions are structured as follows:

- Introduction of conversational digital human with facial emotions.
- A literature review and discussion of prior research in the field.
- A comprehensive explanation of our approach encompasses all the constituent models, components, and achieved outcomes.
- Conducting a user study to evaluate the system's overall performance.

2. Literature Review

At present, technologies and daily life applications are moving towards the digital trend. Embodied Conversational agents are considered a worldwide example that has been around for a very long time for digitalizing human behaviour and interactions. Conversational activities among people involve complex behaviour expressed through speech and gestures, which include facial expressions, hand gestures, head movement and eye gaze. This led to the idea of modelling the human body as well as making it intelligent by combining two systems together: AI and HCI, for many different purposes in different applications and businesses.

An ECA acts as an intelligent entity through conversations by understanding humans and responding back to them using text or voice, both of which are the most commonly widespread techniques. We can find ECAs becoming quite popular through wide industry applications, for example, education, health, business, information retrieval and e-commerce [6]. The main part of an ECA is the chatbot, which is a machine model that processes and simulates the flow of human conversation either in close-domain or open-domain. Chatbots are not necessarily but mostly integrated into interfaces to facilitate user-computer interaction either through text, speech, or both.

Throughout history, chatbots have greatly improved since the date that first enlightened the world with the idea of chatbot, which was in 1950 when Alan Turing suggested the Turing Test, stating the well-known question, “Can machines think?” since then, researchers have begun to compete in introducing different chatbots (Turing, 1950). Chatbots started as text chatbots and are still being used in many industries. Then, speech chatbots were introduced and worked mostly as virtual assistants. 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 [7].

Eliza was the first chatbot developed in 1966, and it aimed to be a psychotherapist. Early Conversational agents such as Eliza depended on predefined simple pattern matching [8]. In 1972, PARRY was introduced to the world and was considered an improvement over ELIZA [9]. After that, in 1995, ALICE, or the Artificial ELinguistic Internet Computer Entity, was developed and was considered to be the most human computer. ALICE introduced AIML, which stands for Artificial Intelligence Markup Language, which is a markup language for manual-defined conversations. Another text-based chatbot is SmarterChild [10], which was developed in 2001 to be used in messenger applications.

Chatterbot software, such as Cleverbot, has been introduced commercially. It uses approaches such as rule-based response generation but is enhanced with techniques for learning new responses. After that came the next step of chatbot creation, which was virtual personal assistants. They became quite well-known and used in daily life tasks by people as they interact with them using voice. The most known examples of virtual assistants are Apple Siri, Amazon Alexa, Microsoft Cortana, IBM Watson, and Google Assistant.

Creating virtual characters necessitates the use of a variety of research skills. Different skills are needed to combine multiple modules within one agent. What plays the most essential role in communication and conversation is the face, which transmits verbal and nonverbal knowledge. Pre-research has introduced new tools like Xface [11], which is an open-source project and a tool to build Embodied Conversational Agents (ECAs) [12].

Text-based and vocal Conversational agents have gained fair popularity in virtual assistant applications. However, various studies, projects, and workgroups have shown that to enhance the reality of interactions with Information and Communications Technology (ICT) systems, it is better to employ embodied interactions. Previous research and projects were done employing a virtual human-like MiraculousLife [13] and CaMeLi [14], which led to the result that the avatar should express more with its facial expressions and give more appropriate reactions. Even though [9] developed successful avatars for virtual social worlds, they lack facial expressions, making the agent seem dull and emotionless.

Due to the rapidly increasing interest in chatbots, particularly post-2016, researchers are striving to develop chatbots that closely emulate human conversational behaviour, resulting in a more human-like interaction. This led to the concept of integrating chatbots into virtual humans. This integration aims to improve human-machine interaction through conversation domains and behaviours, including expressions, acting, and animations. However, sometimes, they are not considered entirely intelligent even though they have bodies and talking heads [15].

Prior studies have not successfully met the essential criteria for enabling a human-like, face-to-face conversation within an intelligent Human-Computer Interaction (HCI) system interface. This goal necessitates establishing the external interface level functions combined with the internal logical level that controls how the system will operate in a behavioural way. In other words, we should take into consideration that a system must give most of the rights that a conversation would consist of between two people, including speech domain, verbal and non-verbal speech acquisition through voice and gestures, and the ability to interrupt while talking and convey emotions.

Although digital humans have been adapted for the entertainment industry, including movies and video games, they are now becoming more prevalent and involved in our daily lives. However, the field still suffers from problems in giving the avatar the concept of a lifelike human. As a result, the field of digital human research continues to improve its realism and overcome its current limitations. Therefore, this research pursues several areas that utilize novel state-of-the-art techniques related to the socialization of digital humans.

The first step towards achieving a real human is its appearance, which falls in the field of computer graphics, and this requires the use of comprehensive deep techniques for remaking a human as a digital one. The choice of tools and techniques is not direct. By following [16], we can understand the difference between the high-quality, expensive approach and the less one in the remaking stage, thus leading to knowing how to relate perception to quality. Despite the focus on external appearance, other research, like [17], focuses on the anatomical side that includes features like skin, muscles, and bones.

Another fact to add is the focus on the balance between the body and the model itself. Alvarado et al [18] did this by proposing a rendering technique to model the deformation of the ground as a result of movement. Nonetheless, a lifelike digital human does not occur through the physical features only but through socialism as well. [18] addressed the aspect of real communication by engaging AI deep learning methods.

With this challenging compilation, it is evident that HCI has advanced so much in recent years and improved in many digital human-related applications and aspects such as lip-syncing and gestures. This inspired us to make use of such novel techniques and try to present a new ECA that will focus on multiple conversational behaviours through verbal speech to non-verbal speech. Our primary objective is to provide a wide speech domain chatbot that will cover a variety of possible topics along with emotion-conveying ability. Our system is considered a combination of skills and novel techniques that are presented by former research, all integrated as one system model manifested as a virtual human in the most straightforward possible way.

3. A Conversational Digital Human with Facial Expressions

Our approach is an embodied conversational agent that can understand and convey emotions, interact with users, and use facial expressions. This requires deploying multiple neural network models. We show our system structure and workflow in Figures 1 and 2, which consist of two parts. In Figure 1, we have three models combined that work simultaneously. These models are speech recognition, response generation, and emotion recognition. We get two inputs when a user talks to the avatar: the utterance that the user said and the face of the user in the video frames while saying that utterance. After that, we convert the input speech to text using the speech recognition model for the avatar to understand what the user says and replies. Then, the resulting text will be fed to the response generator to be analyzed, and a proper reply will be generated as the first output.

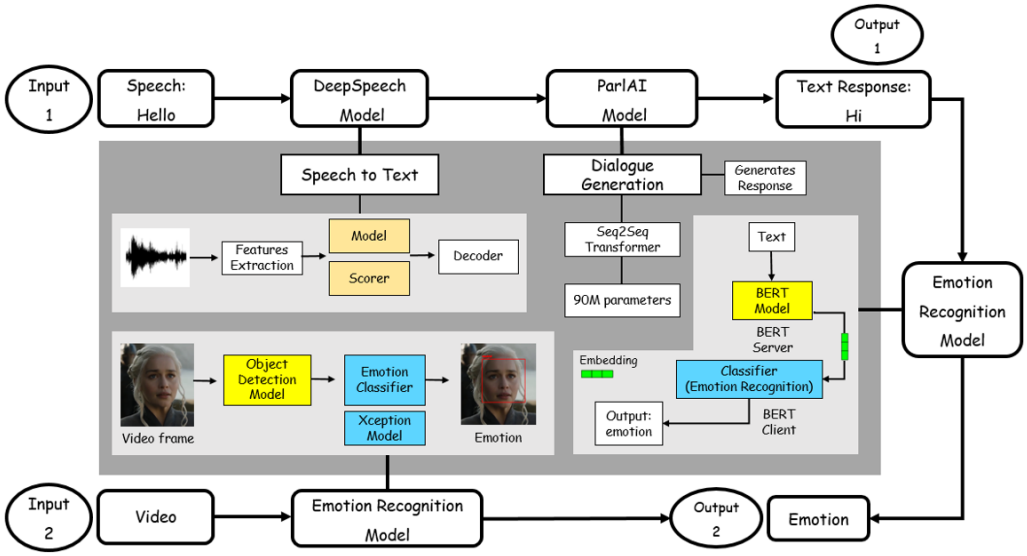

Figure 1. Our system's workflow between the models and with their architectures.

At the same time, our efforts are directed at establishing suitable emotions for the avatar's presentation. Thus, the system will process two inputs: images and text. There will be two emotion outputs, each extracted from different models: the CNN model and the NLP model. CNN will output the first emotion by taking video frames as input, and then the second emotion will be extracted from the response text by the NLP model. Having two similar emotions will be fine; however, we pick the one with a more significant probability if they differ.

In the second part of the system, we work on the Unity Game Engine's side. The two outputs from the previous part are sent to Unity as inputs. As shown in Figure 2, the system synthesizes the text that generates the speech for the response; it also takes the emotion to map the appropriate facial expression through blendshapes and runs the lip-syncing feature in the project. As a result, the avatar speaks back to the user while showing emotion on its face.

3.1. Emotion recognition

Recognizing human emotion is a key component in the study of human-computer interfaces (HCIs) to empathize with people [12,19-20]. Conveying emotion has a crucial contribution to effective human-computer interaction [21].

As the agent should act realistically, we aim to get the accurate emotion to map the appropriate facial expression over the agent's face while conversing with the user. For this purpose, we followed two techniques to combine their results. During the prediction, we compare the probability of predictions done by each predictor to obtain the more accurate one.

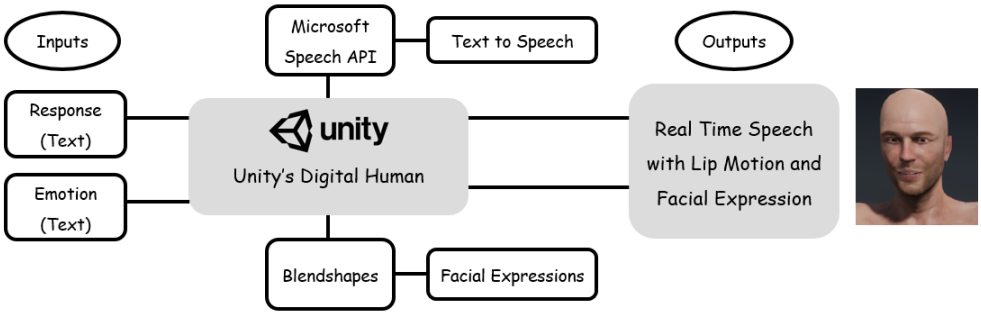

Figure 2. The architecture shows the Unity Game Engine part when combined with result of the deep learning model.

Table 1. Evaluation metrics for text-based emotion recognition model.

F1-Score	0.93
Precision	0.87
Recall	0.83
Accuracy	0.84

3.1.1. Text-based emotion recognition

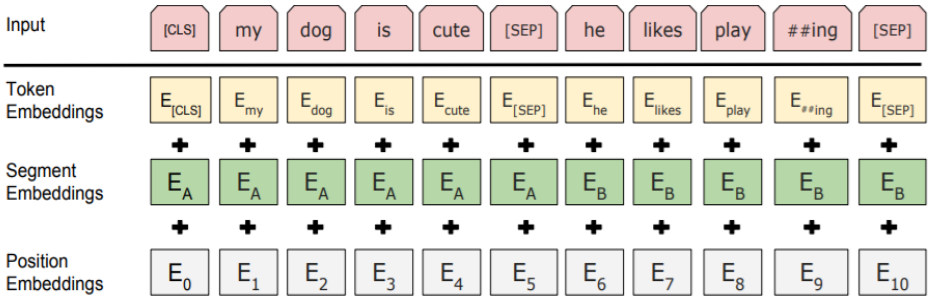

Figure 3. Input representation in BERT includes the sum of embeddings of tokens, segmentation embeddings, and position embeddings, denoted as (E) [20].

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 the text but associates three categories, like "positive" or "negative" or "neutral," for classifying the views regarding a text [22]. Following the same scheme, we get the idea of emotion recognition, which involves analyzing the text and recognizing what type of emotion it implies.

Our agent will respond to the user with a sentence, and to accomplish this, we employ an NLP model that analyzes the response. This analysis aims to detect any hidden emotion within the text, ultimately leading to the generation of an appropriate facial expression corresponding to that emotion.

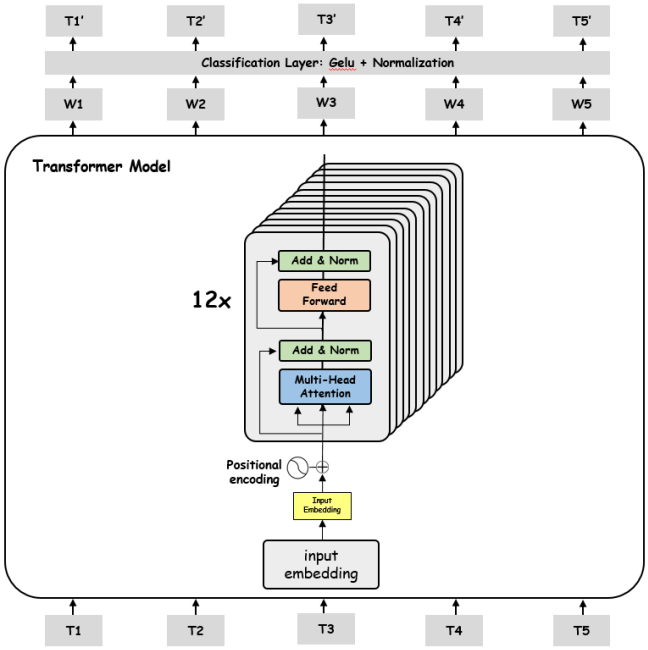

Figure 4. The Transformer (T) based BERT base architecture with twelve encoder blocks.

For this purpose, we fine-tuned Google's BERT pre-trained model that Google AI language researchers recently introduced as a Machine Learning technique built upon state-of-art techniques for NLP tasks [23]. ktrain framework [24] was used as a wrapper for the model to facilitate and accelerate the BERT training process.

The reason for choosing BERT is because it is pre-trained on a massive data corpus, giving it a large knowledge repository and substantial contribution to the NLP community. BERT can be downloaded and fine-tuned on any dataset for NLP tasks. The deep bidirectionality of BERT allows it to learn information from both sides, right and left, as displayed in Figure 3, of the context within the training step, which utilizes adapting it to achieve NLP tasks [25].

BERT architecture is illustrated in Figure 3, which simplifies how the input is processed. In addition, it is essential to mention that BERT has two architectures, BASE and LARGE, that differ in the count of heads in attention modules in encoders, parameters, layers, and units in hidden layers. Our work uses the BERTBASE, which consists of the following values: *parameters*=110M, *encoders*=12, *hidden layers units*=768, *heads in attention modules*=12. In accordance with GLUE benchmarks, BERT performed better than its predecessors. Although BERTLARGE improved BERTBASE's performance, we used the basic one because the dataset is not extremely large, so it does not require a vast network model such as the large BERT version.

BERT is widely used on the basic NLP tasks of classifying a piece of text, which suits our aim in this part of the work as we aim to categorize sentences into emotions classes. In Figure 4, we summarize the architecture of BERTBASE model that we deployed.

We used the following for training the model:

- Ktrain wrapper for Keras,
- Dataset combines three main datasets: daily dialogue, emotion-stimulus, and isear, which includes 7934 inputs for training and 3393 inputs for validation,
- learning rate = 2x10-5, which follows the 1-cycle learning rate policy [26],
- 20 epochs,
- activation function = GeLU.

The resulting output will be one of seven emotions: happy, sad, angry, scared, neutral, surprised, and disgusted. The model achieved 83% accuracy, and the evaluation metrics for the text-based emotion recognition model are shown in Table 1; the evaluation metrics for the real-time video-based emotion recognition model are shown in Table 2, respectively.

3.1.2 Video-based emotion recognition

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 toward a human, the agent will detect the emotion through the expression revealed on the human's face because it is asserted that video-based facial expression is the most informative method for the machine's perception of emotions [27].

This part of the work requires using two models: one for detecting the human's face and the other for detecting the emotion extracted from the human's facial expression. We use the Cascade Classifier for face detection, a machine-learning model for image object detection. For detecting emotions, we operate on a CNN model to learn features and classify emotions.

For this purpose, we fine-tuned one of ImageNet's pre-trained models [28], the Xception model [29], which is pre-trained on the ImageNet database. ImageNet is an image recognition project aiming to classify an image into up to 1000 categories. ImageNet's Xception model consists of 29 layers and is 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 such as VGG16, VGG19, ResNet50, and Inception V3.

We illustrate the workflow of the Xception architecture in Figure 5. For fine-tuning the model, we used the following:

- fer2013 dataset = 35.9k images. 80% is for training, and 20% is for validation.
- Keras library
- 100 training epochs
- Adam optimizer, learning rate = 0.001 and is reduced when validation loss is not improving
- ReLU and Softmax activation functions

It is worth mentioning that we generated real-time data augmentations for the images using Keras's built-in class ImageDataGenerator. The output will fall into one of these emotion categories: "anger," "disgust," "fear," "joy," "sadness," "surprise" and "neutral." Evaluation methods have been applied again to this model; the resulting accuracy is 76%, and other measures are displayed in Table 2.

Table 2. Evaluation metrics for real-time video-based emotion recognition model.

F1-Score	0.75
Precision	0.83
Recall	0.68
Accuracy	0.76

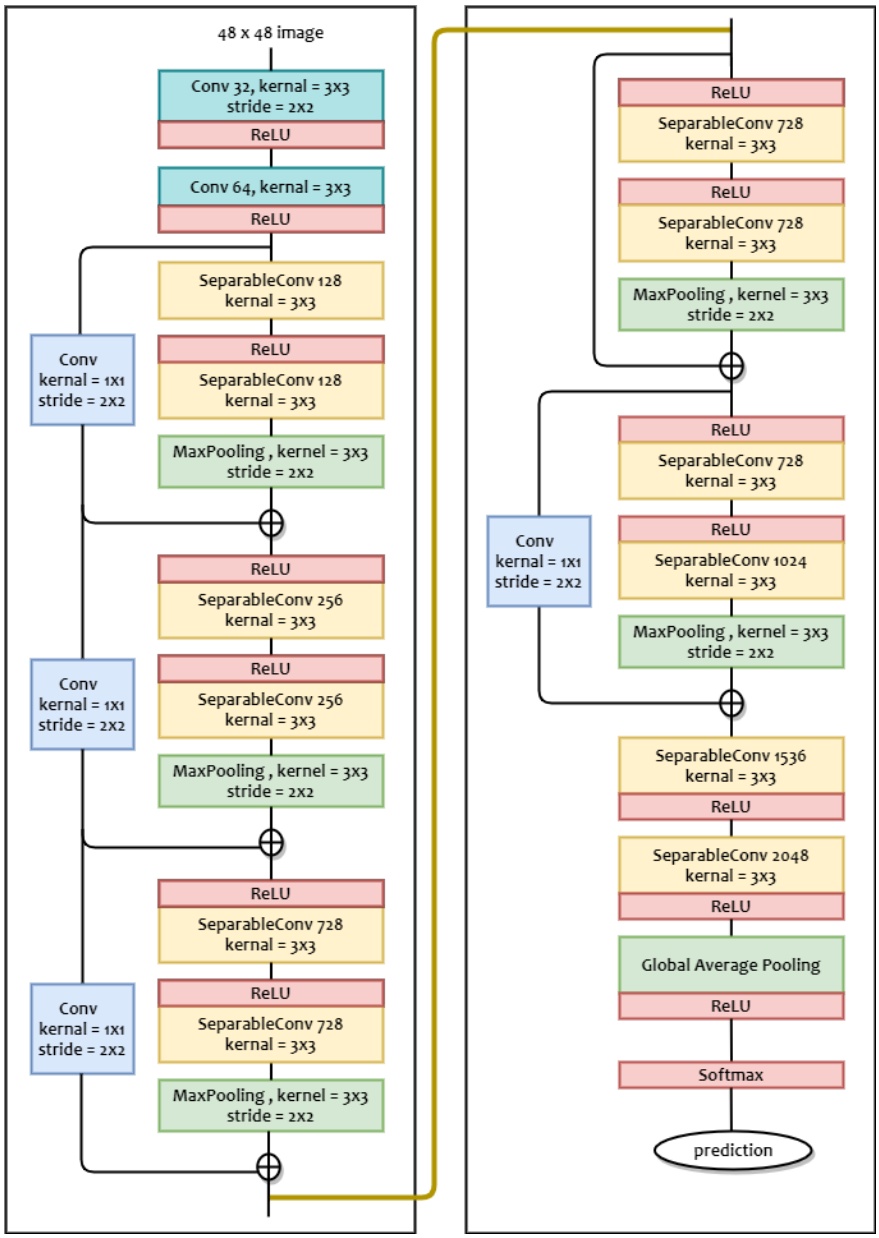

Figure 5. Architecture of Xception network model used for emotion detection from video.

3.2. Dialogue generation

Building an open-domain chatbot is considered a challenge in the ML field, requiring extensive research and a long time. Therefore, we avoid going through the process of building one, and instead, we use the model proposed by [30], which is shared by the ParlAI framework [31]. In this paper, we focus on the Transformer, which allows us to build different chatbots and then fine-tune them. We chose the Generator model, which in its architecture is almost equivalent to the standard seq2seq model introduced by [32], but huger provided three sizes (90M, 2.7B, 9.4B). Rather than retrieving the replicates from a fixed dataset, the generator generates the replicates. We use the 90M model for our project.

The animation of the digital human face involves mapping emotion-relevant facial expressions and applying lip movements to synchronize them with speech when the agent speaks.

3.3. Facial animation

3.3.1. Emotional facial expression

A human's perception, intent, and verbal and nonverbal expressions are expressed effectively through emotional facial expressions; thus, this part of the work is a key major for completing the project. Multiple techniques have been introduced to generate animations, like blendshapes [33-34], bone positions, or a facial action coding system (FACS) [35].

Blenshapes are a well-known technique within digital productions. We can define a blendshape as a geometry deformation for creating looks for the mesh, which means it is originally a group of deformed versions of the mesh blended with the neutral or the regular version of the mesh. Besides being practical for representing various appearances for models, such a technique is also very effective and common in animations and facial expressions.

A blendshape is a linear union of facial targets to create expressive facial expressions and muscle actions [36]. In the virtual human project, we can configure 6 facial expressions for the fundamental emotions according to [37] basic emotions: joy, sadness, anger, fear, surprise, and disgust. We access these emotions through coding to apply proper facial expressions during conversations between the user and the virtual human.

3.3.2. Lip syncing

Lip synchronization means matching lip movement with the speech sound. Lip syncing consists of combining three main stages: facial muscle movements, phonemes, and visemes. We will briefly explain each of them: The phoneme is the smallest unit in a language, like the m sound in Mother and th in thread, whereas visimes 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 turn into visemes.

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

3.4. Speech-to-text

Baidu's DeepSpeech [38] is a deep learning neural network model architecture that manipulates the process of Automatic Speech Recognition and gets an implementation by Mozilla [39], which uses Tensorflow for more straightforward implementation and will be deployed for our system to understand and convert the user's speech to text. The input of the model is the spectrogram, and the output will be a sequence of character probabilities. The architecture includes 3 non-recurrent layers, one bidirectional recurrent layer, and one non-recurrent layer.

Achieving speech-to-text conversion requires the conversion of speech from sound waves to electrical waves. Then, once an analog-to-digital converter converts it to digital data, models can start working on audio to convert it into text.

The hidden Markov Model (HMM) is a widely accepted approach to handling tasks related to speech recognition. 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 SpeechRecognition, which is a library that comprises multiple APIs for speech-related tasks, for instance, the Google speech API. This API supports the default API

key integrated within the library. To choose the Google web speech API, we set the recognize_google() method for the Recognizer class.

3.5. Text-to-speech

To achieve the agent response to humans, we need to synthesize the speech, producing audible output. We preferred using Unity's plugin to synthesize the speech from within Unity instead of deploying a model like Wavenet [40]. To achieve this objective, we utilized the "Microsoft Windows Text-to-Speech API" plugin. This plugin involves creating a wrapper around the Microsoft Speech API in Unity. The underlying technology relies on Windows COM capabilities, introduced with Windows Vista. When the virtual human application in Unity starts running, this wrapper launches the text-to-speech engine, resulting in speech by reading the generated text through the provided function.

3.6. Integrating into unity

Starting with the Unity part, some models need to work in parallel, meaning that running them together at once is required. For instance, when the user is talking, we already know that the user's speech will be converted into text immediately once the user starts finishing a sentence. Also, while the user is talking, we know that emotions appear on that person's face, which means that real-time video capturing has started detecting the emotion from the face. Thus, while the user's speech is being converted with the speech recognition model, the video-based emotion recognition model will also work to extract facial expressions for emotion. To make two processes run together, we used threading in Python. Threads improve performance through parallelism, so we use them to run models together to get outputs simultaneously.

After obtaining the outputs, we utilize communication protocols in networking to send these outputs data with Python-Unity socket communication. Data will be sent to a port through sockets in Python, and then we use the UDP client in Unity to read this data from a socket. We have emotion and response as text data to be sent from Python to C# in Unity. Thus, the C# client will send a connection request to the Python server, and when accepted, the server creates a thread for the client that runs in the background to listen for the requests. For each server client, a socket will be responsible for sending and receiving the data between them. The rest of the output process will be held within Unity C# scripts. All models were combined in one Python virtual environment using Anaconda's command prompt and were called from one Python script, respectively. Using this method, models will keep running and give results in real-time interaction with the agent.

4. User Study

In our study, participants interact with the agent for 10-15 minutes. We first specify tasks and assign each participant a task to carry on the conversation with the agent. Participants wore a Meta Quest 2 Virtual Reality device with a wired connection to start a conversation with the embodied agent. After the conversation, we let the participants give their feedback by filling out a survey consisting of 10 questions or themes that concentrate on several sides of the system to evaluate the overall performance from the users' point of view. Participants followed the Likert scales for filling out the survey, composed of the following scales: strongly disagree, disagree, neutral, agree, and strongly agree, by spending approximately five minutes to finish, concluding the study in twenty minutes total time. The statements and the tasks are shown in Table 3.

We recruited 6 participants (4 females) with an average age of 24 to 30 (Mean:25.1). During the study, they were seated singly in a room while running the system on a PC opposite them. Headphones with microphones were given to the users to talk with the agent. Each user was given a task to talk about with the agent after introducing each. As summarized in Table 3, the first user gets to talk about the first task, which is personal hobbies and interests. Then, the second user talks about movies and video games. The third user talks about holidays and travelling. The fourth user talks about personal issues and problems. The fifth user generally talks about people and what he hates and loves, and finally, the last user asks the agent personal questions and answers in the case being asked.

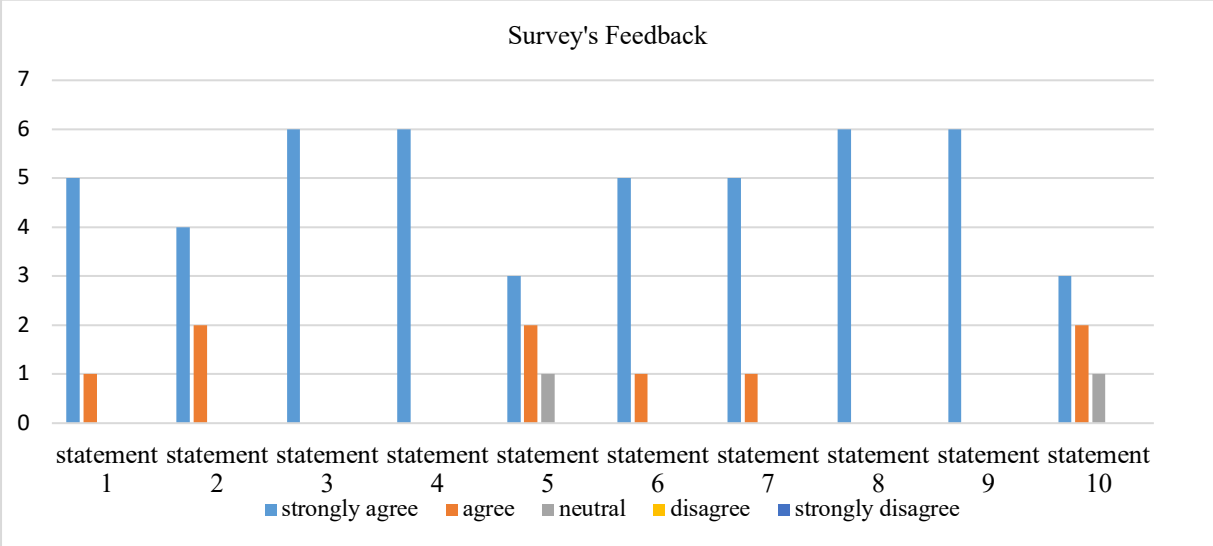

Figure 6. The results of the survey are shown as a graph. For the 1st statement, 5 users strongly disagree, while only one agrees. For the 2nd statement, 4 users strongly disagree, but 2 agree. For both the 3rd and 4th statements, all participants strongly agreed. In the 5th statement, 3 strongly disagreed, 2 agree, and 1 gave a neutral response. For the 6th and 7th statements, 5 strongly agree, and 1 agrees. All participants strongly agreed on the 8th and 9th statements. For the 10th statement, 3 strongly agrees, 2 agree, and 1 is neutral. In all statements, none of the users gave negative feedback by disagreeing or strongly disagreeing.

Table 3. The list of themes or statements asked in the survey (top). The list of tasks given to users to talk about with the agent (bottom).

Statement 1	The agent has human-like interaction.
Statement 2	The agent was persuasive in interaction.
Statement 3	The agent was entertaining and not dull.
Statement 4	The conversation was not boring.
Statement 5	The agent provided quick responses
Statement 6	The agent replied with responses related to the task
Statement 7	The agent provided accurate answers responses.
Statement 8	The agent was expressive in emotions.
Statement 9	The agent reacted to the context with proper emotions.
Statement 10	The overall system, including emotions, conversation, and interaction was realistic.

Task 1	Personal hobbies and interests
Task 2	Movies and video games
Task 3	Holidays and travelling
Task 4	Personal issues and problems
Task 5	People, things to hate and love
Task 6	Personal questions.

5. Results and Discussion

Applying expressive and emotional interaction is the key insight of our work to generate a realistic Conversational agent. To do this, we conducted a user study by arranging conversations between the avatar and multiple users to determine how convincing our model was. As a result, many of the users were satisfied as most of the participants gave positive feedback, proving that our multi-modal avatar can be successful but requires more improvement; we believe that it will make it deployable in many industries for particular goals. Results are summarized in Figure 6, which illustrates the participant's feedback on every theme.

Going through these ratings and starting with the first statement, which refers to how much the digital agent was able to deliver a real interaction, five of the users were fully satisfied with the human-like behaviour by strongly agreeing on it, and more precisely, it was approved in terms of the lip-syncing ability. On the other hand, one person rated the agent with a normal agreement, elaborating that the agent should act differently from one user to another.

Regarding the second statement, which discusses how persuasive the agent was, we received positive feedback from all subjects, but only two of them saw a lack in the persuasion part. These two users commented on the lack of body language during the conversation. Their expectations included hand gestures and head movements to indicate agreement or disagreement when the avatar responded to a question. Despite this, the agent was still approved in the name of entertainment, with all users strongly agreeing. Without a doubt, the key to a successful end-to-end conversation is to have interesting topics, a variety of ideas to talk about, and non-ending responses that prove the agent is not acting boring or dull.

For the fifth statement, we asked the subjects for their opinion on the speed of the agent's responses. One user found it slow during some answers, to which he pointed out that a human's reaction is naturally fast whenever he or she is asked. We also noticed that another user had chosen the "agree" feedback for this, for which he emphasized that there were some moments when he felt that the agent was lagging in the answer as if it didn't understand the question correctly or it took a few seconds to find a correct answer. He also justified this by saying that this is still something that can happen with people because sometimes it takes a moment to think about what to say. For the next ratings, we took the accuracy of the conversation into account and asked our subjects to give their feedback.

Regarding statements 6 and 7, one user gave a less agreeable opinion on both points because she found that one of the answers was not what she expected for her question. This might have happened because the question had a metaphor, and the agent was not able to process the metaphor. However, we believe that this is normal in our case, as the bot is trained on normal everyday conversation data, not deep literature data, and the dataset itself can improve this.

In our study, our focus was emotions. Thus, we proposed two questions on whether the agent was expressive about his feelings and if he was successful in conveying them properly according to the context. All users gave strongly agreed positive feedback. In the final stage of the test, we asked for the final opinion of each participant about the system, including all aspects and whether it was considered too real for them. Though the overall ratings were acceptable as we did not receive any negative feedback, the "agree" and "neuter" votes were still because the system was slow in some reactions and overall lacked gestures.

With this, we conclude our discussion, considering the previous results and realizing that our proposed system will need improvement in terms of movement, whether head or hand gestures. In addition, it should have a whole body to give it more advantages, and it should create its personality like any other natural person has a personality.

5. Conclusion

As machines have become a part of our lives, serving us in every industry, machines need to communicate with humans and enhance their social capabilities. To achieve realistic natural communication and effects on users, the machine or the bot should mimic natural human behaviour and hold humane aspects such as human voice, chatting, emotions, gestures, and expressions. Therefore, this study has demonstrated the potential of combining various advanced AI models to create open-domain conversational chatbots that not only communicate efficiently with users but also provide a friendly virtual avatar and establish realistic interaction with users by conveying emotions through facial expressions. To achieve this, we combined a virtual human with multiple models, each responsible for a task, including speech understanding, speech generation, emotion understanding, emotion generation, and chatting. Although the preliminary study's findings suggest that users generally found their interactions with open-domain conversational virtual avatars to be acceptable, it is evident that additional enhancements are necessary to make these interactions even more lifelike and realistic.

We plan to improve the interaction of the avatar in our future work by providing more realistic interactions with a whole body with more gestures.

References

[1] Robert PH, König A, Amieva H, Andrieu S, Bremond F, Bullock R, Ceccaldi M, Dubois B, Gauthier S, Konigsberg pa, nave s. recommendations for the use of serious games in people with Alzheimer's disease, related disorders and frailty. Frontiers in Aging Neuroscience. 2014; 6:54.

[2] Xiong W, Wu L, Alleva F, Droppo J, Huang X and Stolcke A. The Microsoft 2017 conversational speech recognition system. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 15-20 April 2018; Calgary, AB, Canada. pp. 5934-5938.

[3] Skerry-Ryan RJ, Battenberg E, Xiao Y, Wang Y, Stanton D, Shor J, Weiss R, Clark R, Saurous RA. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In: International Conference on Machine Learning; 10-15 Jul 2018; Stockholm, Sweden. pp. 4693-4702.

[4] Zhang WE, Sheng QZ, Alhazmi A, Li C. Adversarial attacks on deep-learning models in natural language processing: A survey. ACM Transactions on Intelligent Systems and Technology (TIST). 2020 Apr 1;11(3):1-41.

[5] Hyneman W, Itokazu H, Williams L, Zhao X. Human face project. In: ACM Siggraph 2005 Courses; 31 Jul 2005; Los Angeles, CA, USA: pp. 5-es.

[6] Shawar BA, Atwell E. Chatbots: Are they Really Useful? Journal for Language Technology and Computational Linguistics 2007; 22(1):29-49.

[7] Adamopoulou E, Moussiades L. An overview of chatbot technology. In: IFIP international conference on artificial intelligence applications and innovations; 5–7 June 2020; Neos Marmaras, Greece: pp. 373-383.

[8] Griol D, Sanchis A, Molina JM, Callejas Z. Developing enhanced conversational agents for social virtual worlds. Neurocomputing. 2019;354: 27-40.

[9] Weizenbaum, J. ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine. Commun. 1966, 9(1), 36–45.

[10] Molnár G, Szüts Z. The role of chatbots in formal education. In: IEEE 16th International Symposium on Intelligent Systems and Informatics; 13-15 Sep 2018; Subotica, Serbia. pp. 197-202.

[11] Balci K, Not E, Zancanaro M, Pianesi F. Xface open-source project and smil-agent scripting language for creating and animating embodied conversational agents. In: the 15th ACM international conference on multimedia; 25-29 Sep 2007; Augsburg, Germany. pp. 1013-1016.

[12] Aneja D, McDuff D, Shah S. A high-fidelity open embodied avatar with lip syncing and expression capabilities. In: 2019 International Conference on Multimodal Interaction; 14-18 Oct 2019; Suzhou, China. pp. 69-73.

[13] Stainer-Hochgatterer A, Wings-Kölgen C, Cereghetti D, Hanke S, Sandner E. Miraculous-life: An avatar-based virtual support partner to assist daily living. In: ISG 2016 World Conference of Gerontechnolog; 28-30 Sept 2016; Nice, France. pp. 95-96.

[14] Nijdam NA, Konstantas D. The CaMeLi framework—a multimodal virtual companion for older adults. In: Intelligent Systems and Applications (IntelliSys 2016): 21–22 September 2016; London, UK. pp. 196-217.

[15] Don A, Brennan S, Laurel B, Shneiderman B. Anthropomorphism: from ELIZA to Terminator 2. In: the SIGCHI conference on Human Factors in Computing Systems; 1 Jun 1992; San Francisco, CA, USA. pp. 67-70.

[16] Bartl A, Wenninger S, Wolf E, Botsch M, Latoschik ME. Affordable but not cheap: A case study of the effects of two 3D-reconstruction methods of virtual humans. Front. Virtual Real. 2021;2: 694617.

[17] Komaritzan M, Wenninger S and Botsch M. Inside humans: creating a simple layered anatomical model from human surface scans. Front. Virtual Real. 2021; 2:694244.

[18] Regateiro J, Volino M and Hilton A. Deep4D: a compact generative representation for volumetric video. Front. Virtual Real. 2021; 2:739010.

[19] Liu Z, Shan Y, Zhang Z. Expressive expression mapping with ratio images. In: the 28th annual conference on Computer graphics and interactive techniques; 1 Aug 2001; Los Angeles, CA, USA. pp. 271-276.

[20] Queiroz RB, Cohen M, Musse SR. An extensible framework for interactive facial animation with facial expressions, lip synchronization and eye behavior. Computers in Entertainment (CIE). 2010;7(4):1-20.

[21] Lee M, Lee YK, Lim MT, Kang TK. Emotion recognition using convolutional neural network with selected statistical photoplethysmogram features. Applied Sciences. 2020;10(10):3501.

[22] Kharde V, Sonawane P. Sentiment analysis of twitter data: a survey of techniques. International Journal of Computer Applications. 2016, 139(11):5-15.

[23] Kenton JD, Toutanova LK. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT; Jun 2-7 2019; Minneapolis,MN, USA: pp. 4171-4186.

[24] Maiya AS. ktrain: A low-code library for augmented machine learning. The Journal of Machine Learning Research. 2022;23(1):7070-5.

[25] Alammar J. The Illustrated Transformer. https://jalammar.github.io/illustrated-transformer. Accessed 20 April 2021.

[26] Smith LN. A disciplined approach to neural network hyper-parameters: Part 1--learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820. 2018.

[27] Sun Y, Sebe N, Lew MS, Gevers T. Authentic emotion detection in real-time video. In: Computer Vision in Human-Computer Interaction: ECCV 2004 Workshop on HCI; 16 May 2004; Prague, Czech Republic. pp. 94-104.

[28] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition; 20-25 Jun 2009; Miami, FL, USA. pp. 248-255.

[29] Chollet F. Xception: Deep learning with depthwise separable convolutions. In: the IEEE conference on computer vision and pattern recognition; 21-26 Jul 2017; Honolulu, HI, USA. pp. 1251-1258.

[30] Roller S, Dinan E, Goyal N, Ju D, Williamson M, Liu Y, Xu J, Ott M, Shuster K, Smith EM, Boureau YL. Recipes for building an open-domain chatbot. arXiv preprint arXiv:2004.13637. 2020.
[31] Miller AH, Feng W, Fisch A, Lu J, Batra D, Bordes A, Parikh D, Weston J. Parlai: A dialog research software platform. In: the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; 9-11 September 2017; Copenhagen, Denmark. pp. 79-84.
[32] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30.
[33] Smith AP. Muscle-based facial animation using blendshapes in superposition. Doctoral dissertation, Texas A&M University, 2007.
[34] Li T, Bolkart T, Black MJ, Li H, Romero J. Learning a model of facial shape and expression from 4D scans. ACM Trans. Graph. 2017;36(6):194-1.
[35] Prince EB, Martin KB, Messinger DS, Allen M. Facial action coding system. 2015.
[36] Anjyo K. Blendshape Facial Animation, Handbook of Human Motion. Bertram Müller. Switzerland: Springer Cham, 2018; pp. 2145–2155.
[37] Ekman P. An argument for basic emotions. Cognition & Emotion. 1992;6(3-4):169-200.
[38] Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, Prenger R, Satheesh S, Sengupta S, Coates A, Ng AY. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567. 2014.
[39] Mozilla DeepSpeech, https://github.com/mozilla/DeepSpeech. Accessed 1 May 2021.
[40] Oord AV, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499. 2016.