Universitat Politècnica de València

Programa de doctorado en Tecnologías para la Salud y el Bienestar

April 2024

Affective computing framework for social emotion elicitation and recognition using artificial intelligence

Author:

Jose Llanes Jurado

Supervisors:

Dr. Mariano Alcañiz Raya

Dr. Javier Marín Morales

Abstract

The field of affective computing is an area that has emerged with great momentum and is constantly evolving. This field integrates psychophysiology, computer science, biomedical engineering, and artificial intelligence, developing systems capable of inducing and recognizing emotions automatically. Its main focus is the study of human behavior through emotions, which play a fundamental role in actions such as social interaction, decision-making, or memory.

Recently, technological advances have made possible the development of human-machine intelligent systems that were previously unattainable. Elements such as virtual reality and generative models of artificial intelligence are becoming increasingly relevant in the field of affective computing. The combination of these technologies could lead to much more realistic experiments that, along with automatic recognition of physiological responses, could constitute a robust methodology for evoking and identifying emotional states in social dynamic environments.

This thesis focuses on the elicitation and recognition of emotions in virtual reality through the creation of the first virtual human, based on a generative language model, capable of engaging real-time conversations with a human. Various physiological signals have been measured, and signal processing tools have been developed to monitor responses automatically. Based on this, the elicitation and recognition of emotions through machine learning have been evaluated and validated, as well as the recognition of subjects with depressive symptoms.

This work begins with the adaptation of a 2D eye-tracking fixation algorithm to a 3D virtual reality environment. In this context, it is necessary to consider that, in addition to the gaze point, the position of the head within the environment is dynamic, so it must be taken into account to recognize fixations. Since the algorithm depends on a set of parameters, an experimental methodology has been developed, based on the joint optimization of a set of variables, to find the optimal values of these parameters.

The algorithm has been published to be used by the scientific community.

Secondly, the skin conductance signal, which measures sympathetic nervous system activation, is examined. This signal often exhibits motion artifacts that distort the information it conveys. Due to the challenge of automatically detecting and correcting these artifacts, this task has predominantly been performed manually. This thesis proposes the first algorithm for automatic detection and correction of these artifacts, based on deep learning models. A model based on convolutional neural networks with a long short-term memory input layer is developed for artifact detection, improving upon state-of-the-art results. For artifact correction, a regression algorithm is utilized to replace the affected signal. The results of this work demonstrate that the phase decomposition of the manually corrected signal and that which could be performed by an expert do not show significant differences. Conversely, both differ significantly from the original signal with artifacts. The model has been made publicly available for use in systems with automatic processing.

Next, an experimentation for emotion elicitation and detection is conducted. This experimentation is based on a real-time, voice-based conversation with a virtual human. This virtual human has been developed using state-of-the-art artificial intelligence technology, such as generative language models, voice transcription, voice synthesis, and lip synchronization. The experimentation takes place in a semi-immersive virtual reality environment where the virtual human is displayed through life-sized projection. The prototype has been technically validated, and communicative dynamics between the virtual human and the subject have been analyzed. Additionally, the naturalness and realism of the conversations, as well as the elicited emotions in the subjects, have been evaluated.

Finally, the prototype has been used for the recognition of depressive symptoms using eye-tracking and electrodermal activity. For this purpose, an experiment was conducted with 98 subjects, and a methodology for analysis and validation based on machine learning was developed to make predictions. The models achieved a precision of 0.733 and a recognition rate of non-depressive patients of 0.828. Additionally, the recognition of naturalness and realism of conversations, as well as the elicited emotions, has been explored.

The work developed in this thesis presents relevant contributions, not only to the field of affective computing but also, to related areas such as signal processing and human-machine interactions. The tools developed, along with innovative artificial intelligence models, enable experimentation in human social dynamics environments that have never been designed before. The results demonstrate how this work is capable of modeling such relevant information as emotions and depressive symptoms of the person being interacted with. Fields such as psychology, medicine, or education can utilize many of the tools developed in this thesis to provide more information in decision-making or social interaction processes.

Resum

El camp de la computació afectiva és un àmbit que ha sorgit amb gran impuls i està en constant evolució. Aquest camp aconsegueix integrar psicofisiologia, informàtica, enginyeria biomèdica i intel·ligència artificial, desenvolupant sistemes capaços d'induir i reconèixer emocions de manera automàtica. El seu enfocament principal és l'estudi del comportament humà a través de les emocions, les quals juguen un paper fonamental en accions com la interacció social, la presa de decisions o la memòria.

Recentment, els avanços tecnològics han possibilitat el desenvolupament de sistemes intel·ligents humà-màquina que abans no eren factibles. Elements com la realitat virtual i els models generatius d'intel·ligència artificial estan adquirint un paper rellevant en el camp de la computació afectiva. La combinació d'aquestes tecnologies pot donar lloc a experimentacions molt més realistes que, juntament amb el reconeixement automàtic de respostes fisiològiques, podrien constituir una metodologia robusta per evocar i identificar estats emocionals en entorns de dinàmiques socials.

Aquesta tesi es centra en la evocació i reconeixement d'emocions en realitat virtual mitjançant la creació del primer humà virtual, basat en un model de llenguatge generatiu, que permet mantenir converses en temps real amb un humà. S'han mesurat diverses senyals fisiològiques i desenvolupat eines de processament de senyals per monitoritzar les respostes de manera automàtica. A partir d'això, s'ha avaluat i validat no només la evocació i reconeixement d'emocions mitjançant aprenentatge automàtic, sinó també el reconeixement de subjectes amb símptomes depressius.

Aquest treball comença amb l'adaptació d'un algoritme de fixacions d'*eye-tracking* en 2D a un entorn 3D de realitat virtual. En aquest context, és necessari tenir en compte que, a més del lloc de l'impacte de la mirada, la posició del cap dins de l'entorn és dinàmica, pel que cal tindre-la en compte per a reconèixer les fixacions. Donat que l'algoritme depèn d'una sèrie de paràmetres, s'ha realitzat una metodologia experimental, fonamentada en l'optimització conjunta d'un conjunt de variables, per a trobar els valors òptims d'aquests paràmetres. L'algoritme ha sigut publicat per a ser utilitzat per la comunitat científica.

En segon lloc, s'estudia la senyal galvànica de la pell, que mesura l'activació del sistema simpàtic. Aquesta senyal sol presentar artefactes de moviment que distorsionen l'informació que es puga extraure de la mateixa. A causa de la dificultat que suposa la detecció i correcció d'artefactes de forma automàtica, aquesta tasca s'ha realitzat majoritàriament de forma manual. En aquesta tesi es proposa el primer algoritme de detecció i correcció automàtica d'aquests artefactes, basat en models d'aprenentatge profund. Es desenvolupa un model basat en xarxes convolucionals amb una capa d'entrada de *long short-term memory* per a la detecció d'artefactes, millorant els resultats de l'estat de l'art. Per a la correcció d'artefactes s'utilitza un algoritme de regressió que substituiria la senyal afectada. Els resultats d'aquest treball mostren que la descomposició fàsica de la senyal corregida manualment i la descomposició fàsica que podria realitzar un expert, no guarden diferències significatives. En canvi, ambdues sí que les tenen comparades amb la senyal original amb artefactes. El model s'ha fet públic per al seu ús en sistemes amb processament automàtic.

A continuació, es desenvolupa una experimentació per a l'evocació i detecció d'emocions. Aquesta experimentació està basada en una conversa en temps real i per veu amb un humà virtual. Aquest ha sigut desenvolupat utilitzant la tecnologia més avançada d'intel·ligència artificial, com són els models generatius de llenguatge, transcripció de veu, síntesi de veu i sincronització labial. L'experimentació s'ha dut a terme en un entorn de realitat virtual semi-immersiu en el qual l'humà virtual es mostra a través d'una projecció a mida natural. El prototip ha sigut validat tècnicament i s'han analitzat les dinàmiques comunicatives entre l'humà virtual i el subjecte. A més, s'ha avaluat la naturalitat i el realisme de les converses, així com les emocions evocades en els subjectes.

Finalment, s'ha utilitzat el prototip per al reconeixement de símptomes depressius utilitzant biomarcadors d'*eye-tracking* i l'activitat electrodermal. Per a això, s'ha realitzat un experiment amb 98 subjectes i s'ha desenvolupat una metodologia d'anàlisi i validació basada en aprenentatge automàtic per a realitzar prediccions. Els models han assolit una precisió de 0.733 i una taxa de reconeixement de pacients no depressius de 0.828. A més, s'ha explorat el reconeixement de naturalitat i realisme de les converses, així com les emocions evocades.

El treball desenvolupat en aquesta tesi presenta contribucions rellevants, tant per al camp de la computació afectiva com per a altres àrees afins com el processament de senyals i les interaccions humà-màquina. Les eines desenvolupades juntament amb models innovadors d'intel·ligència artificial aconsegueixen realitzar una experimentació en entorns de dinàmiques socials humanes que mai abans s'havia dissenyat. Els resultats mostren com aquest treball és capaç de modelitzar informació tan rellevant com emocions i símptomes depressius de la persona amb la qual es parla. Àrees com la psicologia, la medicina o l'educació poden utilitzar moltes de les eines desenvolupades en aquesta tesi per a proporcionar més informació en la presa de decisions o en la interacció social.

Resumen

El campo de la computación afectiva es un área que ha surgido con gran impulso y está en constante evolución. Este campo logra integrar psicofisiología, informática, ingeniería biomédica e inteligencia artificial, desarrollando sistemas capaces de inducir y reconocer emociones de manera automática. Su enfoque principal es el estudio del comportamiento humano a través de las emociones, las cuales desempeñan un papel fundamental en acciones como la interacción social, la toma de decisiones o la memoria.

Recientemente, los avances tecnológicos han posibilitado el desarrollo de sistemas inteligentes humano-máquina que antes no eran factibles. Elementos como la realidad virtual y los modelos generativos de inteligencia artificial están adquiriendo un papel relevante en el campo de la computación afectiva. La combinación de estas tecnologías puede dar lugar a experimentaciones mucho más realistas que, junto con el reconocimiento automático de respuestas fisiológicas, podrían constituir una metodología robusta para evocar e identificar estados emocionales en entornos de dinámicas sociales.

Esta tesis se centra en la evocación y reconocimiento de emociones en realidad virtual mediante la creación del primer humano virtual, basado en un modelo de lenguaje generativo, que permite mantener conversaciones a tiempo real con un humano. Se han medido diversas señales fisiológicas y desarrollado herramientas de procesamiento de señales para monitorizar las respuestas de manera automática. A partir de esto, se ha evaluado y validado no solo la evocación y reconocimiento de emociones mediante aprendizaje automático, sino también el reconocimiento de sujetos con síntomas depresivos.

Este trabajo comienza con la adaptación de un algoritmo de fijaciones de *eye-tracking* en 2D a un entorno 3D de realidad virtual. En este contexto es necesario considerar que, además del lugar del impacto de la mirada, la posición de la cabeza dentro del entorno es dinámica, por lo que hay que tenerla en cuenta para reconocer las fijaciones. Dado que el algoritmo depende de una serie de parámetros, se ha realizado una metodología experimental, fundamentada en la optimización conjunta de un set de variables, para encontrar los valores óptimos de estos parámetros. El algoritmo ha sido publicado para ser utilizado por la comunidad científica.

En segundo lugar, se estudia la señal galvánica de la piel, que mide la activación del sistema simpático. Esta señal suele presentar artefactos de movimiento que distorsionan la información que se pueda extraer de la misma. Debido a la dificultad que supone la detección y corrección de artefactos de forma automática, esta tarea se ha realizado mayoritariamente de forma manual. En esta tesis se propone el primer algoritmo de detección y corrección automática de estos artefactos, basado en modelos de aprendizaje profundo. Se desarrolla un modelo basado en redes convolucionales con una capa de entrada de *long short-term memory* para la detección de artefactos, mejorando los resultados del estado del arte. Para la corrección de artefactos se utiliza un algoritmo de regresión que reemplazaría la señal afectada. Los resultados de este trabajo muestran que la descomposición fásica de la señal corregida manualmente y la descomposición fásica que podría realizar un experto, no guardan diferencias significativas. Por el contrario, ambas sí la tienen frente a la señal original con artefactos. El modelo se ha hecho público para su uso en sistemas con procesamiento automático.

A continuación, se desarrolla una experimentación para la evocación y detección de emociones. Esta experimentación está basada en una conversación a tiempo real y por voz con un humano virtual. Este ha sido desarrollado utilizando la tecnología más avanzada de inteligencia artificial, como son los modelos generativos de lenguaje, transcripción de voz, sintetización de voz y sincronización labial. La experimentación se ha realizado en un entorno de realidad virtual semi-inmersiva en el que el humano virtual se muestra a través de una proyección a tamaño natural. El prototipo ha sido validado técnicamente y se han analizado las dinámicas comunicativas entre el humano virtual y el sujeto. Además, se ha evaluado la naturalidad y el realismo de las conversaciones, así como las emociones elicitadas en los sujetos.

Por último, se ha utilizado el prototipo para el reconocimiento de síntomas depresivos utilizando bio-marcadores de *eye-tracking* y la actividad electrodermal. Para ello, se ha realizado un experimento con 98 sujetos y desarrollado una metodología de análisis y validación basada en aprendizaje automático para realizar predicciones. Los modelos han alcanzado una precisión de 0.733 y una tasa de reconocimiento de pacientes no depresivos de 0.828. Además, se ha explorado el reconocimiento de naturalidad y realismo de las conversaciones, así como las emociones evocadas.

El trabajo desarrollado en esta tesis presenta contribuciones relevantes, tanto para el campo de la computación afectiva, como para otras áreas afines como el procesamiento de señales y las interacciones humano-máquina. Las herramientas desarrolladas junto con modelos novedosos de la inteligencia artificial consiguen realizar una experimentación, en entornos de dinámicas sociales humanas, nunca antes diseñada. Los resultados muestran como este trabajo es capaz de modelizar información tan relevante como emociones y síntomas depresivos de la persona con la que se habla. Campos como la psicología, medicina o educación pueden utilizar muchas de las herramientas desarrolladas en esta tesis con tal de aportar más información en la toma de decisiones o en la interacción social.

Gracias a Mariano y a Jaime por darme la oportunidad y abrirme las puertas del Leni.

Gracias a Javi por tu esfuerzo y ayuda. Has creído más en mi que yo mismo. Has conseguido sacar lo mejor de mi y juntos hemos hecho cosas que nunca hubiese imaginado.

Gracias a mis compañeros del Leni por hacer del laboratorio mi casa y convertir mi trabajo en un hobby. Me siento muy afortunado de haberos conocido.

Gracias Luu por estar conmigo en tantas batallas y por aprender tanto de ti.

Gracias a mi familia. Gracias Mamá, por ser siempre tan positiva y darme ilusión hasta en los peores momentos, por enseñarme que el esfuerzo en sí ya vale la pena. Gracias a mi Hermana por ser mi ejemplo y mi seguro. Soy muy afortunado de tenerte en mi vida. Y sobretodo, para que esta tesis sea, el éxito de todos los fracasos, gracias Papá.

Gracias por enseñarme la virtud del sacrificio y la pasión por gustarme lo que hago.

Gracias Ana por estar siempre conmigo y por apoyarme. Por ser mi camino. Siempre seré feliz contigo.

Gracias a todos mis amigos y amigas, por cada momento que hemos vivido juntos y que me ha dado tanta vida y ánimos para seguir con todo esto.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

1.1.1 Affective computing

In recent decades, affective computing (AfC) has emerged as a significant area of research, with a focus on analyzing human responses through implicit measures. Introduced by Rosalind Picard, this interdisciplinary field combines elements of psychology, computer science, and biomedical engineering to automate the quantification and recognition of human emotions [1]. Implicit measures have the potential to automatically discern and categorize human emotional states, representing a valuable complement to traditional explicit measures. The capture of emotional insights through implicit measures typically begins with passive sensors that collect data about the user physical state or behavior without actively interpreting the input.

In this context of continuous evolution, the field of AfC is witnessing the integration of diverse technological approaches. The present era showcases various technologies, with a notable emphasis on advances in artificial intelligence (AI), have demonstrated remarkable utility and adaptability to this field. Particularly, the incorporation of algorithms from machine learning (ML) and even deep learning (DL) has become a standard used tools. Additionally, the accessibility of expansive models such as GPT-3, ChatGPT, GPT-4 or StableDiffusion, along with user-friendly libraries like PyTorch and Hugging Face, has taken part in the design and modelization of more complex but sophisticated experimentations. This intersection of AI innovations presents an unprecedented opportunity in the field of AfC, especially in the realms of emotion classification and elicitation. This nascent junction of advanced tools and emotional analysis underscores an emerging area of research.

1.1.2 Emotion modelling

Emotion modelling consists in model human emotions into discrete or continuous emotions. As an example of this task it can be shown the different theorization of emotions showed by Ekman and Russell.

Categorical models classify emotions in discrete categories. This means that each emotion is disconnected from the rest of emotions. The Ekman basic emotion model [2] distinguish between six emotions: anger, surprise, disgust, enjoyment, fear, and sadness. However, categorical models cannot include all the emotions as these models have a concrete number of emotions fixed. This conceptualization could result into a non-optimal emotion detection. It can lead to a mandatory choice identification problem, which is that subjects are likely to distinguish among presented categories rather than to identify an emotion label themselves [3]. Also, the subject could have emotions that are not in the label emotion set. Even if the subject feels a neutral state, it has to pick up one emotion.

On the other hand continuous models recognize emotions quantifying them in dimensions. A common set of dimensions link the various emotional states in this model. They are defined in a two (valence and arousal) or three (valence, arousal, and power) dimensional space. Each emotion occupies a position in this space. One of the most common models is the Russell's circumplex model [4]. This model postulates that every emotion is a linear combination of two affective dimensions: arousal and valence. The arousal dimension delineates the individual psycho-physiological activation related to the emotion, while the valence dimension quantifies the subjectively experienced positivist or negativity of the emotion [5]. This bifurcation results in four distinct regions of the model: High arousal and valence correlates with emotions such as happiness or excitement; high arousal and low valence is indicative of an angry state; low arousal and valence is associated with sadness or depression; and high valence with low arousal is characteristic of a relaxed or contented state.

The majority of studies used the knowledge of these models to classify and predict annotated data, either by using the Ekman's model with six possible emotions or through the Russell's model, quantifying the level of arousal and valence. This task commonly involves the use of different supervised ML algorithms. ML algorithms generally provide more reasonable classification accuracy compared to other approaches, but one challenge in achieving good results in the classification process, is the need to have a sufficiently

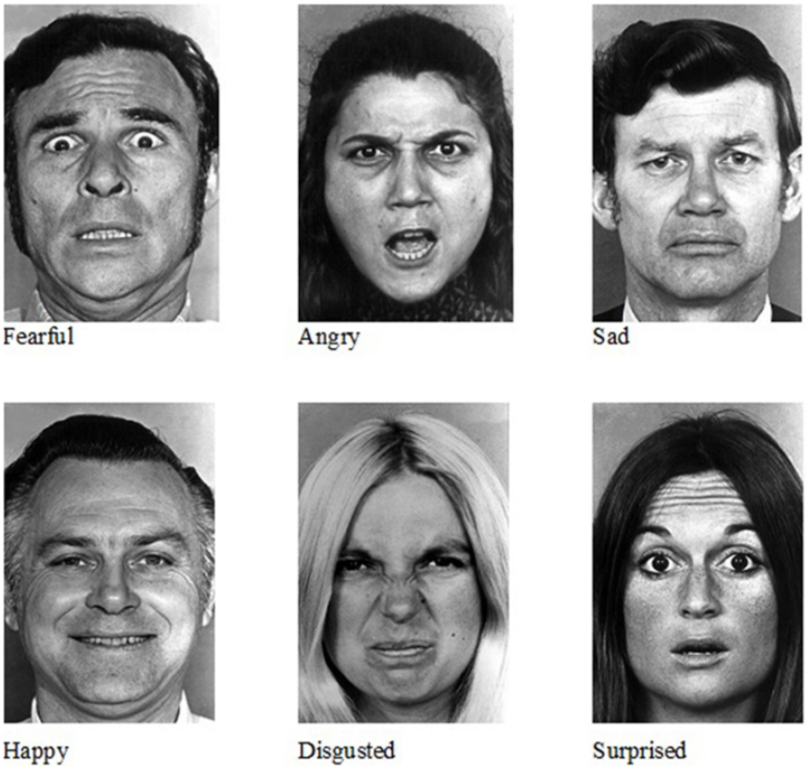


Figure 1.1: Ekman's 6 basic emotions.

large training set [7]. Data is an integral part of the existing approaches in AfC and in most cases it is a challenge to obtain annotated data that is necessary to train ML algorithms. There are many datasets which are public, for the task of classifying emotion types, such as:

- SEMAINE: provides audiovisual recordings between a person and a virtual agent and contains emotion annotations such as angry, happy, fear, disgust, sadness, contempt, and amusement [8].

- DREAMER: provides electroencephalography (EEG) and electrocardiography (ECG) recordings, as well as emotion annotations in terms of valence, arousal, and dominance of people watching film clips [9].

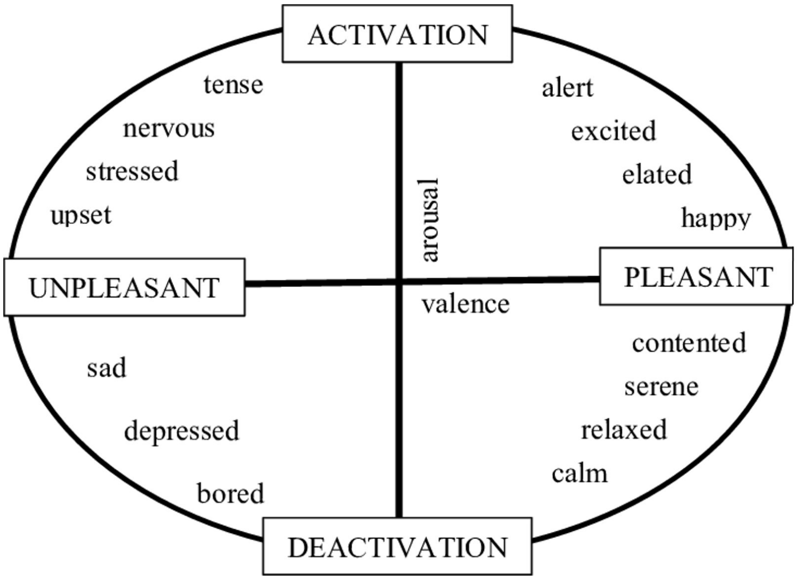


Figure 1.2: Circumplex 2D model theorized by Rusell. Image extracted from [6]

1.1.3 Emotion recognition

Emotion recognition (ER) uses different types of tools to classify emotions through automatic models or human classification in specific tasks. In spite of human emotions variability, it is assumed that there are basic principles, perhaps even basic neural mechanisms that make a particular event "emotional" [10, 11]. To find these principles and their underlying mechanisms, researchers typically study specific emotions, using concrete tasks. As is appropriate, specific experimentations could yield to various types of data, from single neuron firing patterns and activate levels of a concrete brain area. There are different mechanism that could ensemble these data sources and extract patterns from the experimentation in order to classify the different emotions of the subjects during it.

ER studies human emotions through various modalities. Some case of study are the automation of recognizing facial expressions in video, spoken expressions in audio, written expressions in text, and physiological signals measured by wearable devices. The knowledge extracted from these insights is used to assess and study the theoretical models of the multidimensional nature of human emotions.

Emotions influence the activity of the autonomous nervous system (ANS), which, in turn, regulates various physiological parameters. Consequently signals such as heart rate variability (HRV), electrodermal activity (EDA), temperature, and breath patterns are frequently analyzed to recognize emotions. However, the choice of systems for ER depends on the application area and different systems may be suitable for different purposes [12]. It's crucial to acknowledge the complexity of the experimental setup, as it directly impacts the quality of measuring a specific signal.

1.1.3.1 Implicit measures

Traditionally, most theories of human behaviour research have been based on a human mind model that assumes that humans can think about and verbalize accurately their attitudes, emotions and behaviours [13, 14]. Therefore, classical psychological evaluations used self-assessment questionnaires and interviews to quantify subjects responses. However, these explicit measures have been demonstrated to be subjective, as stereotype-based expectations can lead to systematically biased behaviour, given that most individuals are motivated to be, or appear, non-biased [15]. Indeed, research on implicit social cognition suggests that people can act in biased ways without intending to do so. The terms used in questionnaires can also be interpreted differently by respondents, and the outcomes depend on the subjects possessing a wide knowledge of their dispositions, which is not always the case [16]. Recent advances in neuroscience show that most of the brain processes that regulate our emotions, attitudes and behaviours are not conscious and therefore cannot be biased. In contrast to explicit processes, humans cannot verbalize these implicit processes [17].

Implicit measures refer to methods used to assess psychological constructs, such as attitudes, beliefs, or emotions, without relying on direct self-report or conscious introspection [18]. These measures are designed to tap into underlying cognitive processes that individuals might not be fully aware of or might not be able to accurately report through explicit means. Implicit measures are particularly useful when studying attitudes or emotions that individuals might have difficulty expressing or might not even be consciously aware of [19]. Then, implicit measures could capture social behaviour patterns under certain conditions.

Several implicit measuring techniques, to detect affect, have been proposed in recent years [20,21]. They could be divided in two groups: physiological and behavioral signals. Physiological signals encompass multichannel readings originating from the central and ANS, conveying meaningful information regarding actions, responses and feelings [22]. Some examples of their applications in human behaviour research are HRV which has been correlated with arousal changes in vehicle drivers to detect critical points in a route [23], a marker of stress [24] or to correlate it to emotional intelligence as a function of gender [25]. EEG records an electrogram of the spontaneous electrical activity of the brain. EEG is every time more common in AfC experimentations due two reasons. Firstly, EEG data shows effectively cerebral activity information related with human emotional experiences, exploring the neural mechanisms behind them [26]. Secondly, the data recording tool could be integrated properly with the movement of a subject in real world applications. Another common used signal is EDA, which has been used to measure stress, affective arousal and cognitive processing [27–29].

On the other hand, behavioral signals are often spontaneous and may undermine the objectivity of AfC. An example of this type of signals is eye-tracking (ET) which has been used to measure subconscious brain processes that show correlations with information processing in risky decisions [30], empathy [31] and problem solving [32]. Facial expression analysis has been applied to detect emotional responses in e-learning environments [33] and speech emotion recognition has been used to detect depressive disorders [34]. However, people could fake these type of signals such facial expressions or speech to hide real emotions [35, 36].

Numerous studies explore ER not only in specific real-world experiences but also in daily-life habits. Dai et al.'s [37] work investigates the collection of biosignals using a designed bio-sensor. They suggest that ER results can be valuable for emotional health monitoring and serve as key references for clinically diagnosing mental diseases. Boateng [38] focuses on ER within couples, collecting daily data in various situations. Consequently, the study of ER proves to be a versatile field applicable in diverse scenarios and everyday situations. Poria et al. [39] highlight the increasing popularity of ER in conversations, with various AI techniques analyzing conversations ranging from psychological to friendly. Thus, this field is not only intriguing for its potential to indicate the development of mental illnesses but also crucial for understanding how individuals behave in interpersonal interactions.

This work will be focus on two implicit measures, one behavioral and the other physiological, that can be employed using wearable sensors in an ecological setup: ET and EDA.

1.1.3.2 Eye-tracking

ET is the process of measuring either the point of gaze (where the subject is looking) or the motion of an eye relative to the head. The tool used for measuring gaze position and eye movement is called an eye tracker [40]. Eye trackers find applications in research on the visual system, psychology, psycho-linguistics, marketing, and as an input device for human-computer interaction. The configurations of ET setups can vary significantly. There are various technologies that can be employed for ET, such as head-mounted displays (HMD) or ET glasses. ET, especially when combined with VR, is a powerful tool. In immersive VR, the use of HMDs allows not only the presentation of the virtual environment (VE) to the subject but also simultaneous tracking of the gaze position within that environment. The choice of ET tool or the experimental design for ET measurements is crucial in understanding the information provided by this tool. If the eye tracker measures the subject in a VE, it will record the subject's gaze on the virtual content. On the other hand, when using ET glasses, which are more common in non-immersive or semi-immersive environments, the measurements are in the real-world space of the experiment. Extrapolating this data to understand which objects were looked at can involve complex post-processing.

There are various ways to present data from ET research, with static representations and heatmaps being among the basic methods. In static representations, the saccade path is used to connect fixation points above the image. The size of the fixation is typically determined by the fixation duration. On the other hand, heatmaps are representations that aggregate the number of times a region has been visited by the subject. The areas with higher density indicate where users focused their gaze more frequently. Heatmaps are a well-established visualization technique commonly used in ET studies.

The use of ET is highly diverse, finding applications in various fields. In recent studies, ET has been explored as a tool for the early detection of autism spectrum disorder. For instance, the work by Alcañiz et al. [41] utilized ET to identify patterns distinguishing between autistic and control children based on visual attention behaviors in a VE. Another study by Ashraf et al. [42] delved into how ET methodology contributes to training, assessment, and feedback practices in clinical settings. This research emphasized the importance of ET in analyzing the learning curve and providing valuable feedback to users during training. Furthermore, ET is under investigation for car driver distraction detection, as evidenced by the work of Said et al. [43]. Their algorithm assesses instances when the driver is not looking at the road or closes their eyes for an extended period, tested in both virtual and real environments. In summary, ET is a signal that has recently received considerable attention due to its versatility in various experimental setups, demonstrating its potential as a tool for feedback, prevention, or classification.

1.1.3.3 Electrodermal activity

EDA is the property of the human body that causes continuous variation in the electrical characteristics of the skin. Historically, EDA has also been known as skin conductance, galvanic skin response (GSR) or skin conductance response (SCR). The traditional theory of EDA holds that skin resistance varies with the state of sweat glands in the skin. Sweating is controlled by the sympathetic nervous system [44] and skin conductance is an indication of physiological arousal. If the sympathetic branch of the autonomic nervous system is highly aroused, then sweat gland activity also increases, which in turn increases skin conductance. In this way, skin conductance can be a measure of emotional and sympathetic responses [45]. When there are significant changes in EDA activity in response to a stimulus, it is referred to as an event related SCR. These responses, otherwise known as EDA peaks, can provide information about emotional arousal to stimuli. Other peaks in EDA activity that are not related to the presentation of a stimulus are referred to as non stimulus locked SCR.

External factors such as temperature and humidity affect EDA measurements, which can lead to inconsistent results. Internal factors such as medications and hydration can also change EDA measurements, demonstrating inconsistency with the same stimulus level. Lastly, electrodermal responses are delayed from 1 to 5 seconds. These show the complexity of determining the relationship between EDA and sympathetic activity [46]. Often, EDA monitoring is combined with the recording of heart rate, respiratory rate, and blood pressure, because they are all autonomically dependent variables. The skill of the operator may be a significant factor in the successful application of the tool. EDA has a potential potential advantages of low cost and implementability. Also the EDA measure devices are cheaper than other biosignal devices or indeed, the tool to measure EDA signal is included in certain devices setups.

EDA is employed in various types of studies, with a predominant focus on health-related research. This signal can provide valuable feedback for clinical therapies or diagnostics. It is even being used as a measure to model reactive environments in VR. For instance, Liu et al.'s work [47] specifically examines the EDA signal as an indicator of stress levels in patients. They categorize patient stress level as high, medium, and low, achieving a final accuracy of 81.82%. Alcañiz et al. [48] investigates different variables derived from EDA signal processing to recognize patterns of autism in children. Their results demonstrate promising accuracy, ranging between 90.30% and 83.33% depending on the analyzed task. On the other hand, Chiossi et al.'s work [49] introduces a physiologically adaptive system that optimizes a VE based on physiological arousal, specifically the EDA signal. They conduct a simulated social virtual scenario modified based on the detected arousal. Therefore, applications of EDA can be diverse, primarily focused on detecting physiological patterns related to the subject's arousal during an experience. Although this signal is often presented as a complement to other signals like HRV, it can also be studied as a standalone signal, sufficient to reveal certain patterns and trends in the subject.

1.1.3.4 Artificial intelligence in emotion recognition

All the information extracted from the different signals needs to be optimize and study. AI emerges as the best tool to exploit all the ways in which a signal could be modeled. The different techniques that could be applied in the study of a signal become AI in a versatile tool capable to achieve the goal to interconnect signal processing and AfC.

Nowadays there are many experiments that include ML models as a modeling tool, to find patterns in ER. For example, the work of Sharma et al. [50] uses convolution neural networks (CNN) to find the level of engage of a student with a lesson through movements of head and eyes. The work of Tabbaa et al. [51] combines the information of different sources such as ET or GSR to predict the valence and arousal states of the studied subjects. However, ML and DL are not only used as predictive models of the subject's emotional state. These models are also used in the feature extraction process of the different signals. For example, the work of Zemblys et al. [52] employs, for ET, specific tasks such as fixations and saccades classification, or for determining if a blink has occurred, as demonstrated in the study by Medeiros et al. [53].

In conclusion, AI facilitates the exploration of advanced techniques for extracting information from various biosignals, enriching the study of AfC in subjects. However, it emerges not only as a sophisticated statistical tool capable of creating predictive and classification models in AfC, but also, as a tool that enables the development of more complex and versatile experiments.

1.1.4 Emotion elicitation

Understanding behavioral patterns requires not only measuring stimuli but also considering the subject's situational context to identify and activate physiological markers. Foundational theories posit that personality emerges through interactions with situational variables [54, 55]. Personality is viewed as differences in how individuals react to situations rather than context-free individual differences [56]. Variables like extraversion or risk-taking have been found to be weak predictors of behavior in specific situations but strongly correlated with behavioral trends over time [56]. This parallels the evidence supporting the bias of crowds model, where implicit measures are weak predictors of individual behavior in a given situation but strongly associated with aggregated data, as highlighted by Gawronski et al. [57].

The elicitation of emotions is crucial for the ethical and reliable induction of affective states, a key factor in developing systems for detecting, interpreting, and adapting to human affect [58]. Laboratory methods for emotion elicitation can be broadly categorized as active or passive. Active methods involve directly influencing subjects through behavioral manipulation [2], social interaction [59], and dyadic interaction [60]. Passive methods usually present external stimuli as images, sound or video. As to the use of images, the international affective picture system is among the databases most used as an elicitation tool in ER methodologies [61]. It includes over a thousand depictions of people, objects and events, standardized on the basis of valence and arousal [58]. With respect to audio-visual stimuli, many studies have used film to induce arousal and valence [62]. Previous methods have significant limitations, emphasizing the need for high degrees of presence to simulate real-world experiences [63]. Consequently, VR presents a novel approach to emotion elicitation in ER studies by simulating real-world situations in laboratory environments.

1.1.4.1 Virtual reality

The exploration of VR originated in the field of computer graphics and has since expanded into various disciplines [64–67]. In the contemporary landscape, VR-supported video games have gained more popularity compared to the past [68,69], and other fields such as architectural design [70] to education [71], learning, social skills training [72], surgical procedure simulations [73], and support for the elderly. VR facilitates intricate experiments related to navigation studies that would typically require a laboratory setting. Without VR, researchers might have to conduct such experiments directly in the field, potentially with limited control and intervention capabilities.

In the contemporary context, VR emerges as a novel and potent tool for behavioral research in psychological assessment, offering simulated experiences that mimic the real world [74,75]. VR facilitates the simulation and assessment of spatial environments under controlled laboratory conditions, enabling the isolation and modification of variables in a cost-effective and time-efficient manner, which is impractical in real space [76,77].

Moreover, studies have explored VR ability to induce emotions, demonstrating that VE can evoke emotional responses in users [77]. Other research has highlighted the use of immersive VE as tools for emotional induction, generating states of relaxation, anxiety, basic emotions, and examining the influence of users' cultural and technological backgrounds on emotional responses in VR [78–81]. Additionally, studies suggest that emotional content enhances the sense of presence in an immersive VE [82], and when presented with the same content, self-reported emotional intensity is significantly higher in immersive environments than in non-immersive ones [83]. Thus, immersive VEs, whether displaying 360° panoramas or 3D scenarios through a HMD emerge as potent tools for psychological research [84].

Additionally, VR has gained substantial relevance in psychological treatments [85]. De-Juan-Ripoll et al. [86] asserted that VR serves as an invaluable tool for assessing risk-taking profiles and training related skills, with transferability to real world situations. A comprehensive review by Slater et al. [87] presented key evidence of VR applications, highlighting strengths and weaknesses, across diverse research areas such as science, education, training, physical training, and investigations into social phenomena and moral behaviors. The potential of VR extends to fields like travel, meetings, collaboration, industry, news, and entertainment. Moreover, a recent review by Freeman et al. [85], focusing on VR in mental health, underscored its efficacy in both assessing and treating various psychological disorders, including anxiety, schizophrenia, depression, and eating disorders.

The versatility of VR as a stimulus, capable of replacing real world stimuli and recreating otherwise impossible experiences with high realism, has led to its widespread application in research exploring innovative approaches to psychological treatment and training. For instance, VR has been instrumental in addressing issues related to phobias, such as agoraphobia and fear of flying [88]. Additionally, it has been utilized to enhance traditional motor rehabilitation systems [89, 90] by developing games that improve task performance. Specifically, within psychological treatment, VR Exposure Therapy has demonstrated efficacy. This approach enables patients to gradually confront fear stimuli or stressful situations in a controlled and safe environment, allowing therapists to manage psychological and physiological reactions [88].

The different set-ups that could be used to display VR have been continuously evolving during the last years. Nowadays with the increasing use of this technology, more sophisticated technologies could use it easily. Higher or lower degrees of immersion can depend by three types of VR systems provided to the user:

- Non-immersive systems represent the simplest and most cost-effective category of VR applications that utilize desktops to replicate images from the real world.

- Immersive systems offer a fully simulated experience by incorporating various sensory output devices, including HMDs that enhance stereoscopic views through user head movements, as well as audio and haptic devices.

- Semi-immersive systems, like Fish Tank VR, occupy an intermediate position between non-immersive and fully immersive systems. They present a stereoscopic image of a 3D scene on a monitor, utilizing perspective projection aligned with the observer's head position [91]. Advanced immersive systems have demonstrated a closer approximation to reality, creating an illusion of technological non-mediation and instilling a sense of presence in the VE [92]. Moreover, these advanced immersive systems surpass the other two categories by allowing the incorporation of multiple sensory outputs, enabling interactions and actions to be perceived as more authentic [93–95].

Over the past two decades, VR has commonly been presented through desktop PCs or semi-immersive setups like cave assissted virtual environment (CAVE) or Powerwalls [96]. In contemporary applications, there is a growing utilization of HMDs, offering fully immersive systems that effectively isolate users from external stimuli. These HMD-based systems deliver a heightened level of immersion, inducing a more pronounced sense of presence. Presence is defined as the perceptual illusion of non-mediation, giving users the feeling of being present within the virtual scene [97].

1.1.4.2 Virtual humans

Virtual humans (VHs) are computer-simulated entities resembling human characters that commonly engage with humans through computer screens or speakers. Research in this domain revolves around their representation, movement, and behavior, encompassing human-like traits such as speech, gestures, emotions, empathy, and memory. Presently, a VH is essentially a computer program attempting to emulate human characteristics [98]. The core components shaping the body and mind of a VH include embodiment, either in a digital or physical form. The body's purpose is to generate real-time audiovisual content for the VH, enabling live conversational interaction and information reception from users. Conversely, the mind endows the VH with the ability to comprehend natural language, engage in reasoning and creativity, possess memory, and have attributes such as a life history, mood, motivations, and attitudes [98]. The interaction with users through verbal conversation represents one of the most challenging aspects of VHs. They must not only generate coherent, meaningful, and contextualized messages but also retain information and ideas from ongoing conversations. The roots of these interactive tools can be traced back to chatbots and later evolved into conversational agents.

The VH field is a vast research domain comprising numerous research topics such as human movement, facial expression, voice synthetization, memory, communication, interaction with the environment, etc. All in all VH is an emerging field of study that nowadays is increasingly possible with the development of the novel technology.

From Chatbots to Large language models

The first software that allows to interact with natural language were chatbot. They are an informatic system that could establish a conversation with one or more users through different communication channels as voice, text or visual language [99]. However, classical chatbots have a pre-defined sequence of answers to the different possible inputs. It is a bounded system which response is already settled. The use of the chatbots is very diverse. The first operational chatbot is found in *ELIZA* [100] in 1966 and PARRY [101] in 1972. ELIZA and PARRY where used exclusively to simulate a typed conversation with a doctor.

Indeed, these algorithms extract the answer from a database of sentences of doctors. Once the query of the human is done it is compared against the database and the one with highest similitude is the answer of the chatbot. Since there, chatbots have been a very popular field of study. Many different algorithms have been used to improve the communication with a subject trying to overcome the past models. Some of them are *MegaHAL* [102] which is based in Markov's model basing its prediction in a probability distribution choosing between the most likely words for the answer. Chatbots finally evolved when AI algorithms were incorporated to this field.

A large language model (LLM) is a language model characterized by its large size in the number of parameters. As language models, they work by taking an input text and repeatedly predicting the next token or word [103]. Up to 2020, fine tuning was the only way a model could be adapted to be able to accomplish specific communicative tasks. Larger sized models, such as a generative pre-trained transformer (GPT) like GPT-3 [104], however, can be prompt-engineered to achieve similar results. Therefore, these models are allowed to simulate a real conversation with a human. Most of these models allow the introduction of a written context of the conversation, enriching the situation and the way and the messages that the model has to communicate. They have also the capability to remember key points of the conversation, allowing a coherent dialogue with the human. With the actual technology advancement, it is more frequently that a human user could achieve a natural and realistic conversation with a LLM.

To evaluate the naturalness of a conversation there are several tests that could be used, being the Turing test the most famous and exigent one. The Turing test [105], postulated by Alan Turing, was meant as a test for machine intelligence based on whether a human. The test consist whether a human could distinguish, in a conversation with a computer, if it was ruled by a human or not. The machine would pass the test if the human could not distinguish, in different tries, consistently which conversation was ruled by a human and which not. However, with the emergence of VH, this type of test should also be completed with evaluations about the main characteristics of a VH such as interaction or memorization. For example, the 2K Botprize competition [106] is a game bot variant of the Turing test, which replaces chatting with a shooting game environment. The work of Alvarado et al. [107] proposed a test that evaluates aspects of cognitive functioning, associative learning and language acquisition. The work of Pan et al. [108] started to theorize a Turing test for chatbots or VH in VR, and also the challenges that the researches would face against to design certain characteristics such as presence.

VH displays

Technologies which enable the creation of an avatar body for a virtual human, and the ways in which that body can incorporate human, and non-human like senses. Much of the work in this area is being driven by the computer generated imagery (CGI) of the film industry, and the motion capture and animation of the gaming industry. However, for a VH, both imagery and animation need to be generated in real time, and in response to unknown events which becomes this task really challenging [98].

In digital terms, an avatar for a VH (or physical human) can take a number of forms such as:

- A static 2D head-and-shoulders image, as used in many chat applications
- An animated 2D head-and-shoulders or full body image, as used in some customer support applications
- A fully animated 3D character within a game or virtual world.

In creating an avatar for a human or VH, the key areas are:

- Facial rendering which includes mainly lip and eyes movement at rest and for speech animation.
- Body rendering and movement animation at rest and for speech animation. This includes from simple movements as breathing to hand movement during the speech.
- Cloth modeling

Whilst having a high-fidelity digital model of a human face offers one level of problem, having it move to create realistic expressions, and, particularly, to synchronize its mouth with any speech is an even greater challenge. Facial animation includes both, the making of facial expressions (raising eyebrows, scrunching eyes, and smiling) and the movement of the mouth (and neighbouring areas) to match any speech being produced. For the moment, the best tools to achieve the facial animation come from DL models.

In the case of lip-synchronization the main approach of these models is to, given an audio file, predict the position of certain elements around the mouth. This has been investigated in 2D video images and also in 3D avatars. There are a few libraries that accelerate the development of this task. Salsa LipSync Suite [109] provides automated, high quality, language-agnostic, lip-sync approximation for 2D and 3D characters, offering real-time processing of the input audio files to reduce/eliminate timing lag. It is also capable of controlling eye, eyelid, and head movement and performs random emote expressions, essentially providing a realistic face motion for the target 3D characters. The work of [110] use it. At present, Facegood and Nvidia have proposed speech-driven real-time virtual human synthesis schemes respectively. Karras et al. [111] took audio data as input and output of the 3D vertex coordinates of a fixed topology mesh and proposed a 3D face animation driving method based on deep learning and low delay sound. Based on this method, Nvidia implements the audio2face model and embeds it into Omniverse. Like audio2face, there is also the Metahuman creator of the unreal engine [112], Facegood proposes a model Voice2face [113] based on DL, which converts audio into blend shape mixed weight, and combines automatic speech recognition (ASR) and text to speech (TTS) to realize an end-to-end human-computer interaction scheme.

There are different works that have studied how to embody realistically a VH. The work of Döllinger et al. [114] showed how embodying a photo-realistic personalized virtual body affects the awareness of one's internal body signals and how the sense of embodiment is involved in the effects of virtuality and perspective on body awareness.

Unreal engine is another platform for the development of avatars. Unity and Nvidia Omniverse could also develop avatars in 2D and 3D. However, there are not remarkable research that have studied body mechanics in VR for the embodiment of VHs. For the moment this is a research area that has to be studied.

Various studies have indicated that a VH with a human-like appearance tends to convey higher message credibility in advertising contexts compared to those with an anime-like appearance [115]. In the research by Garcia et al. [116], a VH's faces demonstrated high accuracy in ER without engaging in conversation. However, these faces were limited in gesticulation, which prevented the display of various facial expressions associated with emotional states. The work conducted by Karuzaki et al. [117] successfully created a realistic VH using Unity implementation and lips synchronization in VR. Nevertheless, the audio was prerecorded before the conversation, requiring synchronization with the VH's lips beforehand rather than in real-time streaming. Thus, to the best of my knowledge, achieving a VH in VR that can naturally express facial and body movements, along with synchronized real-time audio and lip movement during a conversation, has not yet been accomplished. This highlights that the study of VHs remains an evolving field with countless opportunities for improvement, necessitating the implementation of increasingly realistic methods that faithfully emulate the reality we perceive.

1.1.5 Applications of affective computing in virtual reality

AfC research has mostly used non-immersive 2D images or videos to elicit emotional states. However, immersive VR, which allows researchers to simulate environments in controlled laboratory conditions with high levels of sense of presence and interactivity, is becoming more popular in emotion research [118].

In the realm of AfC research, the predominant approach has revolved around the utilization of non-immersive 2D images or videos as stimuli to evoke emotional responses. Nonetheless, there is a discernible shift towards the adoption of immersive VR. This technology permits researchers to replicate intricate environments within tightly controlled laboratory settings, thereby engendering heightened levels of both sensory engagement and interactivity. The ascendancy of immersive VR is increasingly apparent in the sphere of emotion research, as observed by Marín-Morales et al. [118].

Due to the pronounced sense of presence that VR induces in users, it has been recognized as a potent tool for evoking emotions within laboratory settings. In one of the initial confirmatory studies exploring the effectiveness of immersive VR as an affective medium, Baños et al. [63] demonstrated that both immersion and affective content significantly influence the sense of presence. However, the relevance of immersion was found to be more pronounced in non-emotional environments compared to emotional ones. Subsequent studies further supported the idea that VR can effectively evoke various emotions, including anxiety and relaxation [78], positive valence in obese children engaged in exercise [119], arousal in natural environments such as parks [120], and different moods in social environments featuring avatars [121]. Marín et al. [122] developed four different VEs to predict arousal and valence in each setting, achieving an accuracy of 75.00% and 71.21% for arousal and valence dimensions, respectively. Prabhu et al. [123] designed a VE that utilized biofeedback to alleviate pain and anxiety in patients undergoing total knee arthroplasty, demonstrating promising results during pre and post-operative care. The research by Ontiveros-Hernández et al. [124] underscored the significance of emotion in human activity, particularly in training, using an emotional VR scenario. In summary, VR emerges as a promising tool not only for identifying or inducing specific emotions but also for effective training and learning applications.

1.1.6 Depression

Experiments in AfC not only possess the capability to elicit and identify emotions but also unveil patterns related to deeper emotional states, including those associated with mental health conditions such as depression, anxiety, or schizophrenia. Specifically, depression is a pervasive condition affecting more than 260 million people globally, constituting approximately 3.5% of the global population [125]. Depression influences an individual's thoughts, behavior, feelings, and overall sense of well-being [126], manifesting as a mental state characterized by low mood and aversion to activity [125]. Individuals experiencing depression often exhibit a loss of motivation or interest, along with diminished pleasure or joy from activities that typically bring enjoyment [127]. While it may be a temporary response to life events, such as the loss of a loved one, depression can also serve as a symptom of certain physical diseases and a side effect of medications and medical treatments. Symptoms may encompass sadness, difficulty in concentration, significant changes in appetite, alterations in sleep patterns, feelings of dejection or hopelessness, and, in severe cases, suicidal thoughts.

The continuous advancement of technology enables a much faster and reliable diagnosis of physical diseases. However, this is not the case for mental illnesses, which are also very challenging to diagnose. Moreover, the symptoms that these illnesses may manifest in each patient are highly distinct. Therefore, the implementation of automatic techniques that can achieve diagnosis or identify certain patterns for mental illnesses is increasingly necessary. These tools can facilitate early diagnosis, assess symptom severity, and make appropriate referrals for individuals dealing with conditions like depression. The progress in digital technology provides opportunities for monitoring cognitive and behavioral development, contributing to precision medicine in mental health diagnosis. This involves identifying valid biomarkers and behavioral indicators, enabling the development of personalized preventive and treatment interventions. Such interventions can be tailored to individual characteristics and needs throughout their lifespan. This holistic approach holds promise for enhancing mental health diagnosis and patient care.

1.2 Objectives

The main objective of the thesis is to develop an AfC framework for emotion elicitation and recognition using AI. To do so, we will develop an automatic analysis platform for ET and EDA to perform ER through the use of these biomarkers. In addition, we developed a prototype of VH based on LLM to enhance emotional elicitation simulating social human communicative dynamics. Finally, the developed tools are validated in a use case for ER and depression assessment, completing the thesis objective.

1º Objective. Study and develop an ET fixation identification algorithm in 3D VR to perform automatic feature extraction.

2º Objective. Develop a deep learning model for the automatic identification and correction of artifacts in EDA signal.

3º Objective. Develop and evaluate a VH capable of simulating human communicative dynamics, engaging in voice-based, realistic, and natural real-time conversations. The VH should be designed to express various emotions, with distinct dialogues corresponding to the emotions triggered.

4º Objective. Recognize emotion and depression symptoms during social communicative dynamics with VH.

1.3 Thesis structure

The thesis is structured as follows:

Chapter 2. **Development and calibration of an ET fixation identification algorithm for immersive VR**. This chapter introduces a novel ET algorithm to detect fixations in VR based in I-DT algorithm validated in 2D experimentations. This work also studies the optimum thresholds of the algorithm in terms of different ET features.

Chapter 3. **Automatic Artifact Recognition and Correction for EDA based on DL models**. This chapter shows different DL and ML models that had been studied to identify artifacts in segments of EDA signal. The best model among them is selected, an automatic correction of the identified artifacts was performed and compared against manual expert corrections. Results showed similar performance between the manual and the automatic correction.

Chapter 4. **Developing conversational VHs for social emotion elicitation based on LLMs**. This third chapter presents the whole development and validation of a VH in VR. The VH is allowed to maintain conversations with a human in real time by voice. This chapter explains all the different platforms, applications and AI models used to achieve it. Moreover, different VHs were designed with different emotional states, studying the possibility that the dialogue between the subject and the VH achieves emotion elicitation.

Chapter 5. **Emotion and depression recognition through conversational Virtual Humans**. This chapter studies the emotion and depression recognition through the use of emotional VHs. The work analyzes ET and EDA signals, processing them in order to obtain several features for the identification of ER and depressive patterns. The classification of the targets is done through a ML pipeline and studies which emotional VH could contribute better to a specific target for its recognition.

Chapter 6. **Discussion**. This chapter discuss the results obtained in the different chapters, and the major contribution of the thesis. It also enumerates the possible future lines of work related with the work of this thesis and the framework developed.

Chapter 7. **Conclusion**. Provides an overall summarized conclusion of the work in this thesis.

Finally, the manuscript enumerates the publications and research stages derived from this thesis and provides a list of references.

Chapter 2

Development and calibration of an eye-tracking fixation identification algorithm for immersive virtual reality

Llanes-Jurado, J., Marín-Morales, J., Guixeres, J., & Alcañiz, M. (2020). Development and calibration of an eye-tracking fixation identification algorithm for immersive virtual reality. Sensors, 20(17), 4956. https://doi.org/10.3390/s20174956.

Abstract

Fixation identification is an essential task in the extraction of relevant information from gaze patterns; various algorithms are used in the identification process. However, the thresholds used in the algorithms greatly affect their sensitivity. Moreover, the application of these algorithm to eye-tracking (ET) technologies integrated into head-mounted displays (HMD), where the subject's head position is unrestricted, is still an open issue. Therefore, the adaptation of ET algorithms and their thresholds to immersive virtual reality (VR) frameworks needs to be validated. This study presents the development of a dispersion-threshold identification algorithm applied to data obtained from an ET system integrated into a HMD. Rules-based criteria are proposed to calibrate the thresholds of the algorithm through different features, such as number of fixations and the percentage of points which belong to a fixation. The results show that distance-dispersion

thresholds between 1°-1.6° and time windows between $0.25\ s - 0.4\ s$ are the acceptable range parameters, with 1° and $0.25\ s$ being the optimum. The work presents a calibrated algorithm to be applied in future experiments with ET integrated into HMD, and guidelines for calibrating fixation identification algorithms.

2.1 Introduction

Virtual reality (VR) is a rapidly improving emerging technology [128]. While the gaming industry is taking the lead in the development of VR, it has also found many research applications. VR allows the simulation of experiences which create the sensation of being in the real world [74]. It is very helpful in human-subject-based experiments that are difficult to perform in the real world; it offers environment simulations under controlled laboratory conditions where researchers can efficiently isolate and manipulate features while keeping the other environmental stimuli unchanged [75, 76]. VR not only allows free navigation and real-world type movement [129], it can also evoke similar emotions and cognitive process to physical environments [77]. There are three types of VR, differentiated by the degree of immersion that the technology provides: non-immersive, semi-immersive and immersive [128]. Virtual environments displayed on single-screens, such as desktop PCs, are classified as non-immersive [130]. Powerwall screens, or cave automatic virtual environment (CAVE) technologies, achieve higher degrees of immersion [90, 131]. This technological environment is classified as semi-immersive VR. Immersive virtual environments (IVE), using head-mounted display (HMD) technologies, provide the highest degree of immersion. HMDs isolate the subject from external world stimuli and provide a complete simulated experience [132]. Continuing technical HMD upgrades, such as in resolution and field of view, are increasing researcher's interest in and use of this technology [128, 133]. Technologies such as HTC Vive or Oculus Rift allow six degrees of freedom (DoF) inside the IVE, which is crucial for whole-room VR experiences [134]. Whereas other types of HMD, such as Oculus Go has only three DoF which is the simplest form of user tracking in VR. This is an important difference because increased DoF gives higher sense of presence inside the VR [135].

The development of VR technologies has enhanced research into understanding human behaviour [128]. Moreover, in addition to classic self-assessment, VR can be combined with several implicit measures which model unconscious processes, such as electrodermal activity (EDA), heart rate variability (HRV) [122,136] and eye-tracking (ET). ET is the analysis of eye movements based on corneal reflection and pupil detection.

It is an important source of data for obtaining a complete dataset of features from the analysis of different types of eye movements [137]. ET studies what a subject is looking at [138, 139]. During recent years, developments in ET technology have allowed it to be incorporated into many devices, such as screens, mobile ET glasses and HMDs. ET has a powerful application in VR, as has been shown in many previous studies. For example, Tanriverdi et al. [140] measured the difference in the interaction between eye movements and hand pointing in VR scenarios to assess spatial memory. More recently, Skulmowski et al. [141], using ET, studied the psychological behaviour of subjects in VR, and Juvrud et al. [142] and Clay et al. [143] highlighted and explored ET applications through immersive VR devices.

Eye movement has been studied mainly in two different experimental designs, world-centred and head-centred [144–146]. The principal difference between them is the origin of the coordinate system. In the world-centred design the gaze is directed at the 2D display, while the subjects is positioned in front of a screen, with restricted head movement [147]. This system is used in experimental designs where remote or desktop-integrated eye trackers are used [148]. The head-centred design, on the other hand, measures gaze points within the video recording, which moves as the subject moves his/her head. In this case, the subject can freely move during the experiment; however the origin of the gaze coordinates is in the video display. This has big advantages in real-world experiments, but involves some difficulties in automatically identifying what the subject is looking at. This type of system derives from advanced technologies, such as mobile eye trackers (MET) [148]. However, the eye trackers integrated into HMDs present a new VR-centred framework, where the origin of the coordinate system is the virtual environment. The subject can move freely inside the VR scenario, and the impact point of the gaze is calculated using the intersection between the gaze ray and the polygons of the virtual environment.

One of the most useful methods of modelling gaze-behaviour patterns is through fixation classification. A fixation is defined as a cluster of points where the distance between points is not greater than a certain value and its temporal interval is longer than a certain time. Intuitively, it has been interpreted as a group of points where a subject has focused his/her gaze [139]. There has been extensive discussion in the literature about the minimum time and dispersion distance to define a fixation. It has commonly been considered that the minimum fixation time has to be above 0.1 s [149]. The minimum time depends on the task being performed by the subject. For tasks such as reading and visual search, the minimum fixation time stipulated is 0.225 s and 0.275 s, respectively. For tasks where eye-hand coordination is required, the mean fixation time has been established at 0.4 s [139]. In summary, mean fixation time has been established as between 0.15 s to 0.65 s [138]. The dispersion angle of fixations is not as yet defined but they are normally fixed below 2° [139]. To perform fixation classification analysis, there are three main types of spatial criteria algorithms; velocity-based, dispersion-based and area-based [150]. Velocity-based algorithms use the eye's velocity information assuming that fixation points have low velocities and saccades points higher ones. One of the most popular algorithms is called velocity-threshold identification (I-VT). Dispersion-based algorithms emphasize the spatial distance between points using at the same time temporal and spatial information. They are based on the idea that spatial distance is lower in fixations than in saccades. The robustness and accuracy achieved is better than the two other types of dispersion based algorithms [150]. A representative algorithm from this type is the dispersion-threshold identification (I-DT). Finally, area-based algorithms identify a group of points which are inside an area of interest (AoI). All three types of ET fixation identification algorithms need a set of thresholds and spatial and temporal information to classify eye movements.

These fixation identification algorithms are mainly applied in world-centred and head-centred experiments. Very little work has been done on fixation classification in the VR-centred paradigm [143,151]. In world- and head-centred paradigms only 2D gaze vectors have been studied; however, for VR-centred, ET provides two 3D vectors, gaze and head position [147, 152]. How to apply both sets of vectors, in order to study eye movements, is an underaddressed challenge. Duchowski et al. [151], to obtain visual angle, proposed a solution where head position is averaged for every set of points that can be included inside a fixation. While this solution is a valuable contribution, the specification of the optimum parameters for a concrete ET fixation classification algorithm in VR-centred design remains an open question. Duchowski et al. [151] introduced the parameters of the 3D implemented algorithm, manually aligning the gaze-interaction points of subjects with the environmental targets displayed. On the other hand, it is not believed that the gaze acquisition that derives from the VR engine has an influence on eye-signal frequency, as this is influenced by the velocity of the renderization. In a later world-centred study, Bobić et al. [153] examined the number of predicted and real saccades in a guided task, using an I-VT algorithm. Whereas other studies, such as [145, 154, 155], did not report the exploration of different sets of parameters for fixation classification tasks. It is critical to identify in the related literature the best set of parameters for the algorithm, because very different results can be obtained, and interpretations made, depending on the parameters used [149, 156, 157]. Moreover, the methodology that should be used to identify this optimum region is still an open issue. There is no clear way to infer which is the optimum set of parameters for an ET fixation classification algorithm. For example, Blignaut [157] researched the optimum dispersion threshold for a I-DT algorithm in a free world-centred task, examining different features, and obtained an optimum region between 0.7° and 1.3° for radius threshold, using a time window of 0.1 s. There is still no consensus of how to achieve the optimum parameters for an ET fixation classification algorithm.

This study proposes a new methodology to calibrate a VR-centred fixation classification algorithm using ET integrated into an HMD. A guided experiment was designed to study the fixation identification of an I-DT algorithm applied to an IVE. While there is no ground truth that identifies the optimum parameters for any particular feature [157], four different features were examined in this task. A set of rules was established for each, with the aim of reaching an agreement between the features as a criterion to achieve the most suitable parameters. A final set of optimum thresholds is proposed for use with the I-DT algorithm in future research.

2.2 Materials and Methods

2.2.1 Participants

A group of 57 healthy volunteers (27 females and 30 males), with normal or corrected-to-normal vision, was recruited to participate in the experiment. The mean age of the group was 25.36 (SD = 4.97). The inclusion criteria were as follows: age between 18 and 36 years; Spanish nationality; not having any previous VR experience. All methods and experimental protocols were performed in accordance with the guidelines and regulations of the local ethics committee of the Polytechnic University of Valencia.

2.2.2 Virtual Environment and Data Collection

The virtual environment was displayed through an HTC Vive Pro Eye, an HMD with an integrated ET system (see Figure 2.1), offering a field of view of 110°. The scene is displayed with a resolution of 1440 × 1600 pixels per eye, with a refresh rate of 90 Hz. The set-up includes HTC Wireless Adapter, and an HTC base station covering a $6 \times 6\, m^2$ area. The ET data were obtained from the Unity VR through the ET SDK (SRanipal), with a maximum frequency of 120 Hz and an accuracy of $0.5° - 1.1°$. The computer used was an Intel Core i7-770 CPU 3.60 GHz with an Nvidia GeForce GTX 1070.

To perform the study an immersive 3D scenario, using the Unity 3D [1] platform, was developed. This features a room modelled by an occlusive Cube Map, which is a Unity object that captures all the possible stimuli on the object's surface. This type of object, ensures that all the projections of the eye-tracker rays impact against an element in the scene to provide continuous feedback of the subject's gaze.

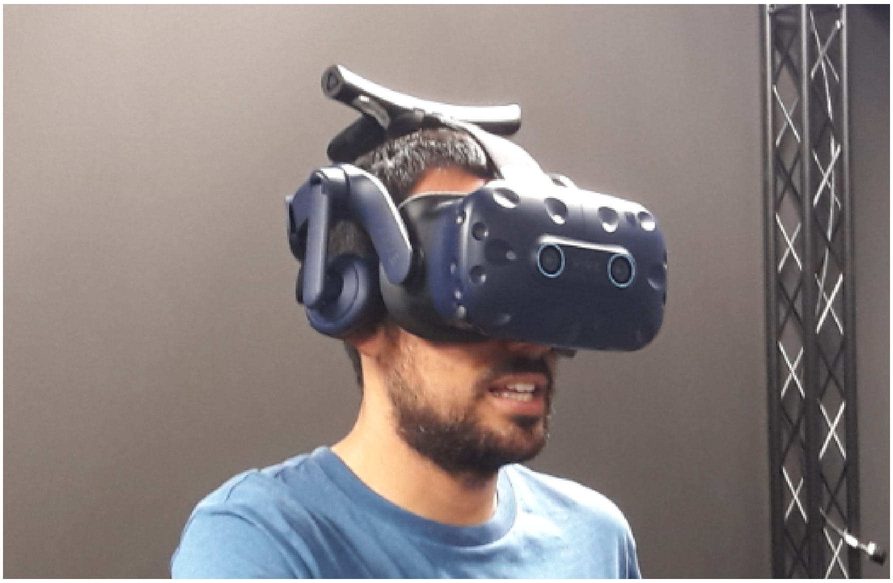
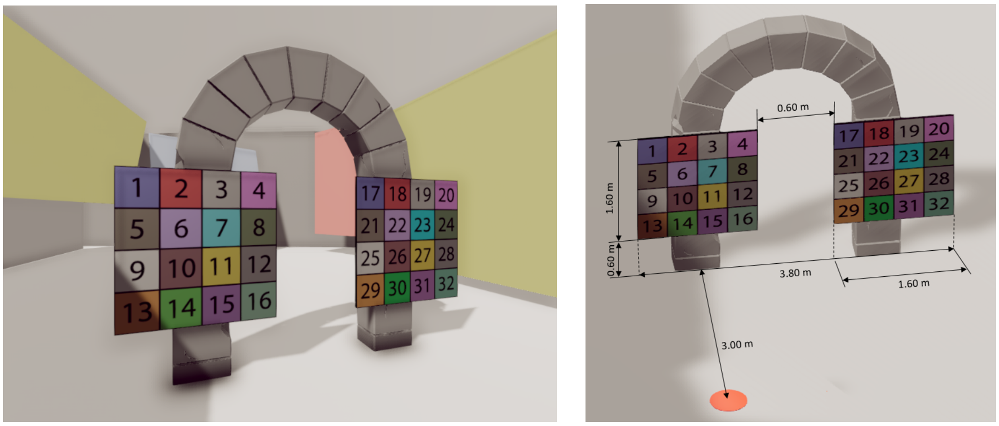


Figure 2.1: Example of a subject using the HTC Vive Pro Eye for the development of the experiment.

The room includes two large similarly-sized panels. Each panel displays a matrix of 4 × 4 numbers, where every square is identified by a sequence from 1 to 16 in the first panel, and 17 to 32 in the second. Each square includes a background colour to ensure contrast between the cells and focus the subject's attention (see Figure 2.2(a)). The initial location of the viewer is above a marked orange point in the scene, in front of one of the panels (Figure 2.2(b)). The location of this orange point was established to provide frontal gaze to one panel (from $-14.93°$ to $14.93°$), and diagonal gazes in the other one (from $25.02°$ to $45.00°$), to ensure that the subject moved his/her head during the experiment.

[1] https://unity.com/



(a) Virtual Scenario from subject perspective.

(b) Perspective view of the scenario.

Figure 2.2: Virtual scenario screenshots. The orange dot (b) designates the position of the subject.

Several squares were lit following a pre-determined sequence designed to evoke many different fixations. The subjects were asked to look at the illuminated squares during the task. The sequence was the same for all subjects. It had been created randomly with the following guidelines: It had to begin in the front panel and explore its four diagonals and its centre. Next, the subject had to look at the furthest and nearest points of the second panel (on the right). Finally, from a certain square the sequence changed, such that the squares lit alternated between the panels, first the left, then the right, etc. The resulting sequence was 1, 16, 4, 13, 6, 11, 7, 10, 17, 32, 22, 10, 20, 5 and 30. The subjects were asked to freely explore the environment for 1 minute to adapt to it. After that, every square was lit for 3 seconds following the predetermined sequence, the total time of the guided task being 45 seconds. Every lit square was defined as an AoI.

The raw eye-tracking data included the 3D position of the impact of the gaze ray in the environment, and the 3D head position in the virtual space. This data is the input of the I-DT algorithm. The gaze point includes the coefficient, for each eye, of the probability that an eyelid movement constitutes a blink, where 1 means completely closed, and 0 open. Points above 0.75 in either eye were considered as blink points and removed [152]. This represented 0.69% of the total raw data. Moreover, the virtual environment exports a file, which recorded when a specific square was lit (e.g. square 1 ; time $0\ s - 3\ s$). This file was used to synchronize the gaze data with the illuminated sequence protocol. Only data that were, in terms of time, between the first and last lit squares were taken into account.

2.2.3 Fixation identification algorithm

The algorithm implemented is an adaptation of an I-DT algorithm. A previous study suggests the use of this algorithm due to its robustness and accuracy in the fixation identification task and its low number of parameters (dispersion and time threshold) [150]. Moreover the I-DT algorithm has been used in many previous ET parametrical studies [156,157]. In accordance with the VR-centred paradigm, the algorithm considers 3D points which are intersections of the gaze rays with virtual objects whose origin is the 3D head position. For fixation identification, head position was averaged every time that a point was added as a candidate to be a part of the fixation. Averaging the head position of the subject will take into account the free 3D movement of the subject inside the VR. Then, each beam considered as a part of a fixation will have its origin in an averaged head position [151]. It is important to note that the methodology followed or the algorithm used are not dependent on the dimensionality of the experiment. Both could be used with 2D data, however this work examines a VR-centred experiment designed in 3D. To measure the distance between a set of points, dispersion distance (DD) [156] was used. DD measures the angular distance between the pairs of points that are candidates to be a part of a fixation. The algorithm records a set of consecutive points with time differences smaller than a specific value (line 2). The highest distance in the group has to be less than the dispersion threshold value (line 5 and 6 of Algorithm 1) to consider the set of points as a potential fixation. The dispersion threshold and time window are parameters which have to be set initially to the I-DT algorithm. Both are essential for the fixation classification task. In addition, the algorithm we present applied, as an innovation, a frequency threshold below which possible fixation points were discounted (line 3 and 6 of Algorithm 1). This ensured that the algorithm did not use gaze data recorded at frequencies that do not facilitate the detection of fast eye movements. The temporal decrease of the data collection frequency could be provoked by an increment in the graphic renderization requirements of the GPU environment or saturation of the computing capacity, which needs to be taken into account in a VR-centred framework. Since the raw ET data was obtained using the ET SDK (SRanpial) through a Unity script, the frequency of the data depends on the processing velocity of the graphic engine. Therefore, although the ET device works at 120 Hz, acquisition will be lower if the Unity rendering frequency is lower, which is highly dependent on the GPU of the computer used and the complexity of the environment. Frequency variation during the experiment was analysed. To ensure the quality of the fixation classification, the frequency threshold was set at 30 Hz, the lowest frequency the literature uses to study ET data [152]. The pseudocode of the algorithm can be seen as follows.

```
Algorithm 1: Dispersion Algorithm
Data: Data($t, x_g, y_g, z_g, h_x, h_y, h_z$)
Parameters: Dispersion Threshold, Time Threshold, Frequency Threshold
Start
while New gaze point is available do
    Initialize window to cover the Time Threshold
    if Points Frequency > Frequency Threshold then
        Computation of $\theta$ for each pair of points
        if $\theta_{max}$ <= Dispersion Threshold then
            while $\theta_{max}$ <= Dispersion Threshold and
            Points Frequency > Frequency Threshold do
                Add samples
            end
            All samples except the last one are classified as a fixation.
            Remove all this window samples.
        else
            Remove first sample
        end
    else
        Remove first sample
    end
end
```

Where t is the time, x_g, y_g, z_g are gaze coordinates and h_x, h_y, h_z are head position points. The dispersion angle θ is obtained from the scalar product between two vectors, eq. (2.1)

$$\cos\theta_{ij} = \frac{\vec{d_i} \cdot \vec{d_j}}{|\vec{d_i}||\vec{d_j}|} \text{ , with } \quad \vec{d_n} = \vec{g}_n - \vec{\overline{h}} \text{ .} \tag{2.1}$$

The sub-indexes i and j are two arbitrary points and d_n is the final end-point of the subject's gaze n, the origin of which is the average head position in each component $\vec{\overline{h}} = (\overline{h}_x, \overline{h}_y, \overline{h}_z)$ [151].

2.2.4 Calibration criteria

To identify suitable parameters for the I-DT algorithm a parametrical analysis examined different features. No feature exists that ensures the optimum set of parameters, or any ground truth, that can quantify how good are the parameters used by the ET algorithm [157]. An agreement between four different features was used as an appropriate criteria to evaluate the algorithm's optimum parameters, dispersion threshold and time window. The features used in the study are averaged between all the subjects. They are discussed over the next paragraphs; a set of requirements was established. The four features used were, number of fixations, percentage of points classified as a part of a fixation, mean fixation time and percentage of fixations inside AoIs. These features were examined in terms of the dispersion thresholds and time windows in a grid search [156, 157]. The objective is to find a set of points that simultaneously satisfy the conditions imposed for each feature. This set of points would be the optimum to use for an I-DT algorithm in a VR-centred experiment. The grid used to calibrate the algorithm started at 0° and went to 2.5°, in steps by of 0.1°, and time windows from 0.1 s to 0.5 s by intervals of 0.05 s. It is important to note that the first three features can be computed for guided and free tasks, whereas the fourth can be computed only for guided task protocols where specifics AoIs are defined. This calibration method used the first three features to obtain the optimum calibration, and the fourth, a specific feature which depends on the type of study, to specify and obtain the final calibration results.

- **Number of fixations** This measures the average number of fixations per subject during the task. For small dispersion thresholds the growth of the feature increases from zero until a maximum. After this maximum, the feature decreases until one single fixation for a high dispersion threshold and any time window value is obtained. The parameters to be selected must all exceed the maximum number of fixations due to the high instability in this region [157].

- **Percentage of points classified inside a fixation** measures the amount of points classified as part of a fixation. This feature is linked to the number of fixations. In the first step, with a small change in the dispersion threshold, the percentage of points increases exponentially. However, this increasing tendency changes when a maximum number of fixations is reached, which produces an elbow point in the feature [157]. After this point, the growth of the curve becomes smoother until it reaches 100% of the points included as part of a fixation. This feature helps identify a lower-limit for the dispersion threshold. Moreover, this feature has to be as high as possible to classify more points as fixation points [157]. Following this condition, this feature determines a single point and not a region of points.

- **Mean fixation time** This measures the average fixation time per subject. As has been seen in previous studies, this feature follows a linear relation with the dispersion threshold of the I-DT algorithm [156, 157] in 2-D world-centred experiments. The mean time fixation increases proportionally as this parameter increases. This is due to the fact that the more points there are inside a fixation, increases in the dispersion threshold involve increases in fixation time. Parameters which involve mean fixation times above certain mean times cannot be considered as optimum. This condition defines an upper-limit for the dispersion threshold and time window. In this work, the predefined maximum mean time is established at 1.5 s. Despite the fact that this fixation time is higher than the upper limit established in the literature (0.65 s [138]), it is close to the results obtained by [151] in a IVE, that is, 1.9 s mean fixation time.

- **Percentage of fixations in AoI.** This measures the percentage of fixations with centres inside AoIs. A similar feature was used by [151,153] to calibrate algorithms. In the present work, the majority of the fixation centres were found to be inside the defined AoIs. The percentage is obtained when the AoI is lit. This measure not only provides spatial information about where the fixation centre is located, it also provides temporal information, because it only records gaze points inside an AoI when it is lit. With the variation of the parameters of the algorithm, different numbers of points are classified as part of fixations and the positions of the centres of the fixations also change. The feature starts with the highest value (close to 100%). While the dispersion threshold is increased, more points are therefore part of the same fixation. In consequence, the center of the fixation is displaced in order to be in the average position of all the fixation points. These displacements cause the centre of the fixation to be placed randomly in the environment for high dispersion threshold values instead of being centred around an AoI. This evolution tends to induce a decreasement in the percentage of fixations inside AoIs. However, we hypothesise that during this decrease there is a stable region where the variation of the parameters does not affect the percentage of fixations inside the AoI. This region represents the set of parameters that better model visual attention, as the values of this feature are unaffected by small changes in the parameters. The search of this stable region requires two previous steps. First, a simple moving average

(SMA) of three points is used on the signal in order to smooth it and eliminate noise. After that, the first derivative of the feature is computed. Stability is considered to be established when the variation in points is below 2%. Table 2.1 summarises the criteria and the features used to calibrate the ET algorithm.

Table 2.1 shows the criteria and the features used to calibrate an ET algorithm.

Table 2.1: Calibration criteria of the features.

Measure	**Criterium**
Number of fixations	After the maximum fixation number
Percentage of points classified inside a fixation	After elbow point and as high as possible
Mean fixation time	Lower than certain predefined time
Percentage of fixations inside AoI	Stable region

Based on the criteria in Table 2.1, the strategy followed in the calibration process was: (1) Computation of the features which do not depend on the definition of an AoI as being constituted by number of fixations, percentage of points classified inside a fixation and mean fixation time, (2) to compute the optimum value based on maximizing the percentage of points which belong to a fixation, (3) Step 1 is recomputed by including the percentage of fixations inside AoIs. (4) Step 2 is repeated. Therefore, to obtain the optimum set of parameters, steps 1 and 2 take account only of the free-task related features, while 3 and 4 take account of these features and the guided task related feature.

2.3 Results

2.3.1 Frequency analysis

The evolution of the frequency of the eye-tracking data acquisition averaged by all the subjects during the experiment is shown at Figure 2.3 including mean and standard deviations. The fluctuation of the frequency is mainly between $44-46$ Hz where the mean frequency is 44.95 Hz. However, the ET data recording frequency of one subject was below 10 Hz between $34.5-35$ s This anomalous frequency caused the high-frequency variation shown in Figure 2.3. The acquisition of the data mostly complies with the minimum accepted frequency established in 30 Hz.

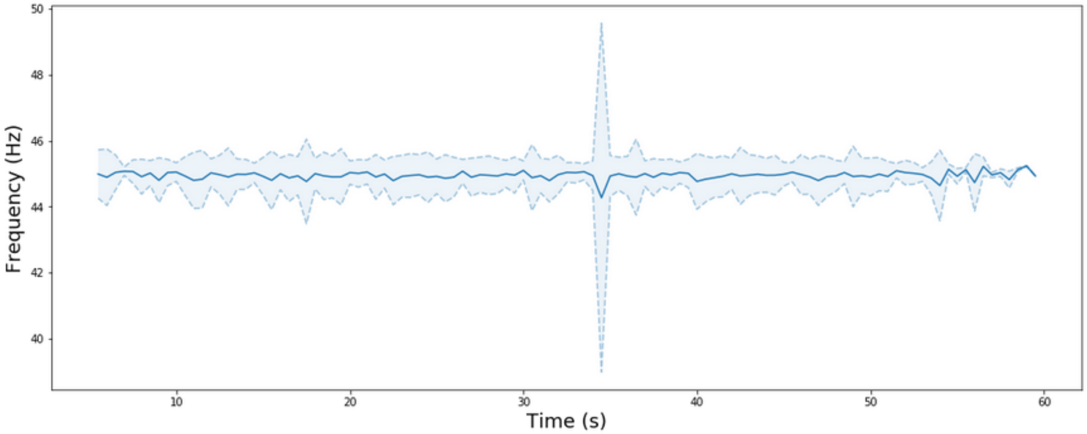


Figure 2.3: Temporal evolution of the ET data frequency. The blue line is the average frequency by subject for each second. The discontinuous line indicates the standard deviation above and below the mean frequency.

2.3.2 Algorithm calibration

Figure 2.4 shows that the number of fixations strongly depends on both parameters. Two different regions can be distinguished in this feature. The first region is defined from 0° until the maximum number of fixations, between $0.25° - 0.6°$. This region shows an increase in the number of fixations until the maximum is reached. After that, the features decreases smoothly as both parameters are higher. The highest number of fixations is achieved with the minimum time window (0.1 ms) and with a dispersion threshold of 0.25°. However, this maximum becomes smoother with greater time windows.

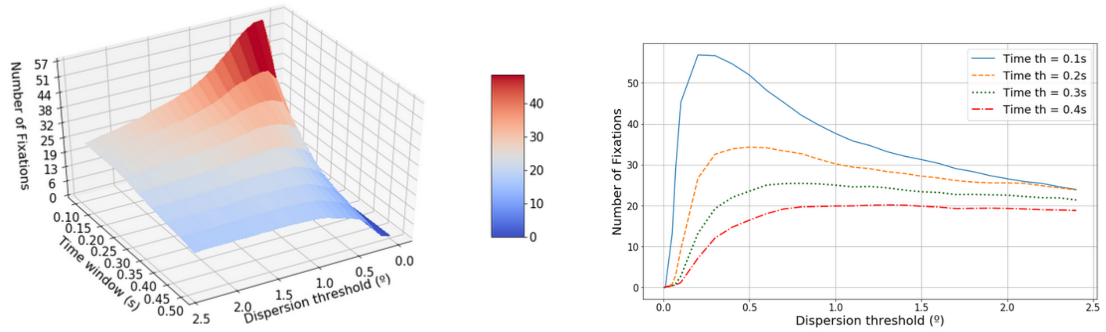


Figure 2.4: The left-hand graphic shows the evolution of the average number of fixations in terms of the dispersion threshold and the time window. The right-hand graphic shows the projected dispersion threshold.

The evolution of the percentage of points classified as fixation points is shown in Figure 2.5. Between 0 ° and 0.5 ° the curve grows exponentially until an elbow point. This elbow point accords with the maximum number of fixations (Figure 2.4) for each time window [157]. However, this inflexion point is more difficult to detect as the time window increases. The percentage of points inside a fixation decreases as the time window lengthens, but increases with the increment in the dispersion threshold.

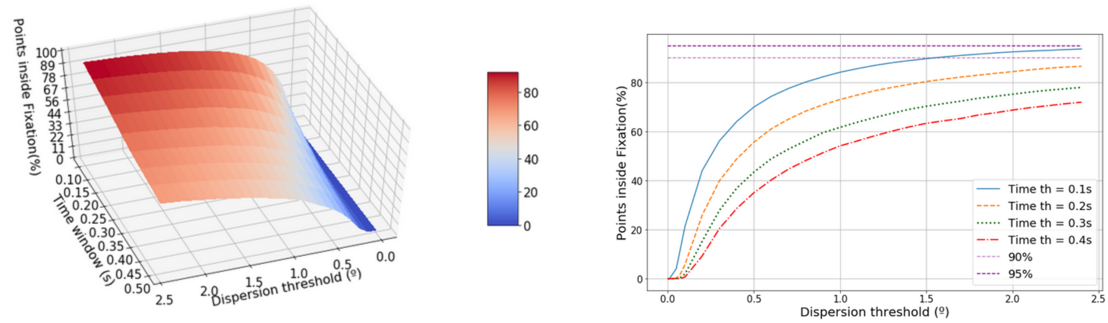


Figure 2.5: The left-hand graphic shows the evolution of the percentage of points classified as a part of a fixation in terms of the dispersion threshold and the time window. The right-hand graphic shows the projected dispersion threshold.

Figure 2.6 shows that the mean fixation time depends linearly on the dispersion threshold and does not depend on the time window for the plotted region. The higher is the dispersion threshold the higher is the mean time fixation. A mean time of 1.5 s is achieved for a dispersion threshold of 1.5°, whereas for a value of 0.5° a mean time of 0.65 s is achieved.

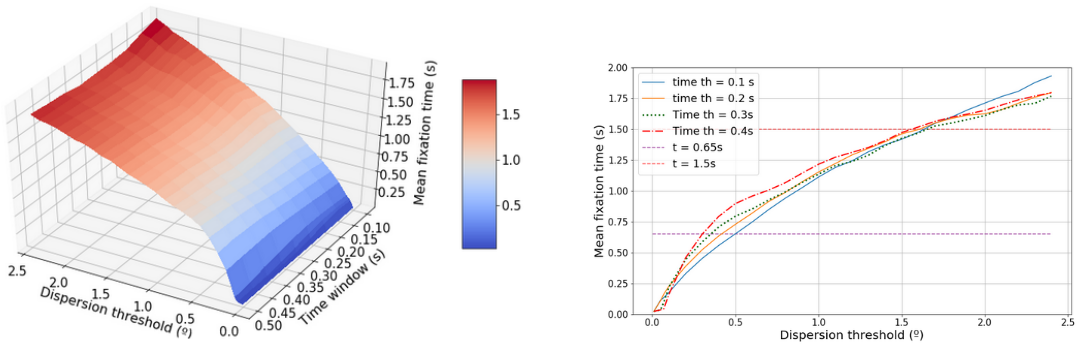


Figure 2.6: The left-hand graphic shows the evolution of the mean fixation time in terms of the dispersion threshold and the time window. The right-hand graphic shows the projected dispersion threshold.

Taking these results into account and applying the rules established in Table 2.1, the optimum set of parameters can be inferred from these computed features. The rules derived from the features, number of fixations, percentage of points classified as a fixation, and mean fixation time, specify a region of parameters (shown in Table 2.2). Moreover, based on our optimization strategy, the percentage of points classified as fixations should be as high as possible; this is achieved for a time window of 0.1 s and a dispersion threshold of 1.6°, with a value of 90.43%.

Table 2.2: Optimum parameters using the criteria of number of fixations and mean time fixation.

Parameters	**Values**							
Dispersion th (°)	0.3 – 1.6	0.5 – 1.5	0.6 – 1.6	0.8 – 1.7	0.9 – 1.6	1.4 – 1.6	1.4 – 1.5	1.4 – 1.5
Time window (s)	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45

The feature percentage of points inside AoIs was computed and added to the results. As can be seen in Figure 2.7, the percentage of points inside AoIs has a high degree of dependency on the time window value, as it varies from 76 % with 0.1 s to 91 % with 0.4 s, maintaining the dispersion threshold in 1°. A flat region is demonstrated for time windows above 0.2 s – 0.50 s (Figure 2.7) and a dispersion threshold between 0.5° – 1.2°. This stability is not seen for time windows from 0.1 s to 0.2 s which have a decreasing trend. As the dispersion threshold parameter increases from points higher than 1.3° – 1.5°, the feature value decreases to zero.

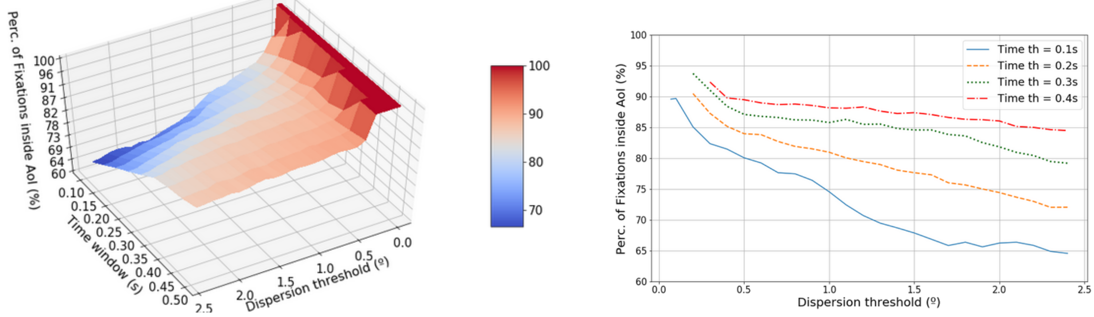


Figure 2.7: The left-hand graphic shows the evolution of the percentage of fixations inside AoI in terms of the dispersion threshold and the time window. The right-hand graphic shows the projected dispersion threshold.

The established criterion of stability for the feature percentage of fixations inside AoIs was added to the results of Table 2.2. A final acceptable set of points that fulfil all the conditions of Table 2.1 was obtained, and is shown at Figure 2.8.

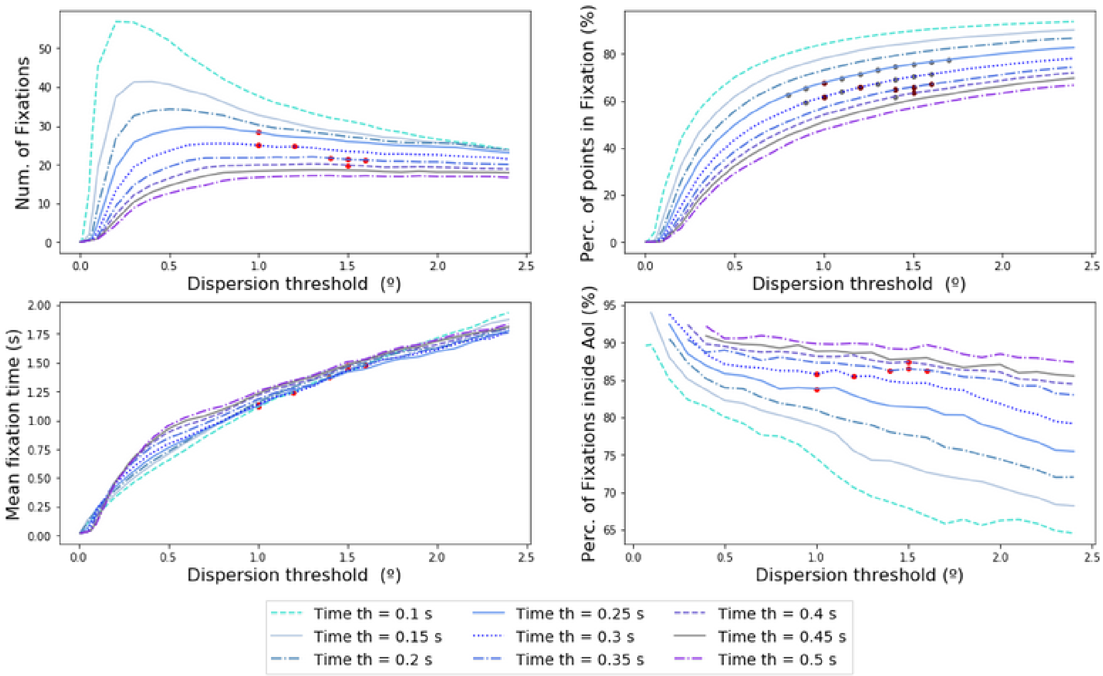


Figure 2.8: The four different features used to calibrate an I-DT algorithm with the points that fulfill all the conditions established (red points) in terms of the dispersion threshold for different time windows.

Table 2.3 shows the values of the red points of Figure 2.8. Only four time windows fulfil the conditions imposed by the Table 2.1. Points where the time window is below 0.25 s are excluded due to high variability in the feature percentage of fixations inside the AoIs. Longer time windows, such as 0.45 s and 0.5 s, are not optimum because the elbow point has not been reached yet. Dispersion points below 1° and higher than 1.6° were also discarded, due to the instability of the percentage of fixations inside these AoIs and due to their large mean fixation time values. The results showed a positive correlation between the optimum dispersion points and the optimum time window: when the time window is larger, the dispersion threshold also increases. Following the criteria that the percentage of points classified as fixations should be as high as possible, the highest value found was 67.82%, for a time window of 0.25 s and a dispersion threshold of 1°.

Table 2.3: Optimum parameters using the criteria from number of fixations, mean time fixation and percentage of fixations inside AoI.

Parameters	**Values**			
Dispersion th (°)	1	1 and 1.2	1.4 – 1.6	1.5
Time window (s)	0.25	0.3	0.35	0.4

2.4 Discussion

The purpose of this study is to develop an I-DT fixation algorithm to be used in a VR-centred experiment and obtain optimum thresholds for the algorithm in a 3D IVE. The results can be discussed on four levels: (1) the novelty of the algorithm used, (2) the optimum thresholds obtained; (3) comparisons with previous studies; (4) the calibration procedure used.

The present study experimentally validated the use of head movements using an I-DT fixation identification algorithm, with an integrated ET sensor in a new generation HMD, the HTC Vive Pro Eye. Moreover, we introduced a new frequency threshold. As can be seen in Figure 2.3, the frequency of the data in an IVE is not always continuous, and it can suffer from high variability due to the IVE renderization; this needs to be considered in VR-centred eye-tracking research. The algorithm presented is robust in the face of these fluctuations, and rejects possible fixation classifications for points below a certain frequency threshold. In addition, as the maximum frequency of the ET device used was 120 Hz, and acquisition in this relatively simple environment was 44.95 Hz, the complexity of the environment needs to be taken into account as it strongly depends on the GPU hardware used. On the other hand, the I-DT algorithm used in this work is not dependent on the dimensionality of the experiment. It could be used to analyze fixation classification in 2D data. This leads to a more general approach to studying ET data.

The strategy followed to calibrate the I-DT algorithm in this research, was to obtain an agreement between the features computed, following the rules established in Table 2.1. The analysis showed the dependency of the number, duration and position of each fixation in terms of the I-DT parameters. Based on the criteria established for every feature computed, a final set of parameters for the calibration of the I-DT algorithm in an IVE was achieved. Thresholds between 1° and 1.6° for dispersion and time windows between 0.25 s and 0.4 s are the optimum, where the point that best fits the fixed criteria is 1° and 0.25 s. It is interesting that the optimum points found followed a proportional relation between the time window and the dispersion threshold. To select concrete values inside this optimum region, the feature that has to be boosted has to be known [157]. For example, with values of 1° and 0.3 s, more fixations will be obtained, but the percentage of points classified as being inside fixations will be less than the classification achieved for parameter points 1.2° and 0.3 s. The results showed a parametrical region that might be considered for future studies where the I-DT fixation classification algorithm is used in an IVE. To the best of the authors' knowledge, these results present the first calibration of a fixation algorithm in VR.

There are many similarities between the results obtained in this work and previous studies. The features examined in this work, obtained using a VR-centred system, follow similar shapes and trends to features computed in world-centred scenarios [156, 157], with different numerical values. Blignaut (2009) [157] showed that the number of fixations achieved is higher than 50 for optimum parameters in an experiment of 15 s duration. The mean fixation time achieved in the present study was 0.25 s for a dispersion threshold of 1.45°. This mean time accords with the mean fixation time established in the literature for world-centred experiments, which is between 0.15 s and 0.65 s [138]. However, Shic et al. (2008) [156] obtained a mean fixation time of 1 s for a dispersion threshold of 1°. On the other hand, VR-centred experiments, such as that of Duchowski et al. (2002) [151], identified a total number of fixations between $15 - 30$ for a 44 s experiment and a mean fixation time of 1.9 s. In the present study, the maximum number of fixations obtained was 28 for the optimum set of parameters, for an experiment of less than 60 s duration. The mean fixation time obtained from the optimum parameters was between 1 s and 1.5 s. The mean fixation time is higher in this study but this might be due to two reasons. The first might be related to the experimental methodology, where the subjects have to look at a fixed AoI for at least 3 s. The second could be that ET in VR manifests some different characteristics in comparison to 2D schema, which accords with the results obtained by [151]. Also, the linear relationship between the mean fixation time and the dispersion threshold agrees with the findings of [156,157]. Thus, the results suggested that the number of fixations is lower, and mean fixation time is higher, in VR-centred than world-centred experiments.

The present work used a guided experiment which indicated where the subject had to look every time. However the VR design could introduce bias in the subject's gaze due the used colors and numeration, which is a limitation of the present study. Previous studies, such as [151, 153], also used this type guided experiment to calibrate their own ET algorithms. This methodology allows the researcher to identify what the subjects have been looking at, at any particular time. This is a key point because this methodology allows the researcher to know and define AoIs and record different features to enable spatial and temporal comparisons between subjects. This comparison is not possible in a free-experimentation task because this does not allow a comparative analysis of AoIs between subjects. In the present study the features which can be computed in guided and free tasks, such as number of fixations, gave us preliminary knowledge of the region of agreement of the parameters. The results show dispersion thresholds of 0.3° to 1.6°, and time windows from 0.1 s to 0.45 s, where the most suitable set of parameters was found for 1.6° and 0.1 s. This parametrical region accords with the results achieved by Blignaut (2009) [157], an experiment developed with a free ET task. When the feature specifically assigned to measure the guided task, the percentage of fixations inside the AoI, was used, the optimum parametrical region was reduced, which provided more accurate results. The results obtained reduced the time window parameter to the interval 0.25 s to 0.4 s, and the dispersion threshold to $1° - 1.6°$, the optimum point being 0.25 s and 1°. This feature complements and specifies the parametrical research. For this reason, a guided experiment provides features appropriate to calibrate a specific ET algorithm. On the other part, the exposure time of the subject to the VR is low, in order to avoid ocular fatigue. However, it would be interesting to evaluate how this fatigue could affect to the results of the optimum parameters in longer experimentations.

2.5 Conclusion

In conclusion, the present study has demonstrated the implications of using an I-DT algorithm in an IVE, which includes some key points as the head movement of the subjects, previously presented by [151]. Moreover, a new frequency threshold is introduced in order to avoid the variation of the frequency which comes from the IVE. The algorithm presented is robust in the face of these fluctuations, and rejects possible fixation classifications for points below a certain frequency threshold. Four different features were used, as no definitive feature exists for modelling visual attention through ET fixation identification algorithms. Different conditions were established for each feature, the optimum thresholds being those that simultaneously accomplish all the conditions of the features. This ends up with a set of parameters which are between 1° and 1.6° for dispersion and time windows between 0.25 s and 0.4 s. However, the point that best fits the fixed criteria is 1° and 0.25 s. We presented a simple case of a guided task, as the experiment did not attempt navigation. Future work will be needed to address how a navigation task influences the calibration. This work established rules to calibrate an ET algorithm; these could be modified based on experiments undertaken and the objectives of the studies. The recent technological developments in VR and ET open a huge new research field combining both technologies. It could mean a breakthrough in the analysis of human behaviour in controlled experimental set-ups using immersive VR. The analysis presented, the type of methodology, and the criteria used in this work, provide a useful guide for future research in the use of ET for fixation classification studies in IVE. Furthermore, this research presents a novel I-DT algorithm adapted to a VR-centred ET paradigm, and some innovations, such as frequency acquisition and the use of 3D head movement for the I-DT algorithm. The algorithm was calibrated and obtained a final set of optimized thresholds, which might be used as a tool in future research analysing gaze patterns in HMDs.

Supplementary Material

The I-DT algorithm in VR-centred system is implemented in the following link
https://github.com/ASAPLableni/VR-centred_I-DT_algorithm

Chapter 3

Automatic Artifact Recognition and Correction for Electrodermal Activity based on LSTM-CNN models

Llanes-Jurado, J., Carrasco-Ribelles, L. A., Alcañiz, M., Soria-Olivas, E., & Marín-Morales, J. (2023). Automatic artifact recognition and correction for electrodermal activity based on LSTM-CNN models. Expert Systems with Applications, 230, 120581. https://doi.org/10.1016/j.eswa.2023.120581.

Abstract

Researchers increasingly use electrodermal activity (EDA) to assess emotional states, developing novel applications that include disorder recognition, adaptive therapy, and mental health monitoring systems. However, movement can produce major artifacts that affect EDA signals, especially in uncontrolled environments where users can freely walk and move their hands. This work develops a fully automatic pipeline for recognizing and correcting motion EDA artifacts, exploring the suitability of long short-term memory (LSTM) and convolutional neural networks (CNN). First, we constructed the EDABE dataset, collecting 74 h EDA signals from 43 subjects collected during an immersive virtual reality (VR) task and manually corrected by two experts to provide a ground truth. The LSTM-1D CNN model produces the best performance recognizing 72% of artifacts

with 88% accuracy, outperforming two state-of-the-art methods in sensitivity, AUC and kappa, in the test set. Subsequently, we developed a polynomial regression model to correct the detected artifacts automatically. Evaluation of the complete pipeline demonstrates that the automatically and manually corrected signals do not present differences in the phasic components, supporting their use in place of expert manual correction. In addition, the EDABE dataset represents the first public benchmark to compare the performance of EDA correction models. This work provides a pipeline to automatically correct EDA artifacts that can be used in uncontrolled conditions. This tool will allow to development of intelligent devices that recognize human emotional states without human intervention.

3.1 Introduction

Electrodermal activity (EDA) is a non-stationary signal that indicates electrical potential via the sweat glands on the surface of the skin [158]. EDA represents a quantitative functional measure of sudomotor activity and, therefore, an objective assessment of emotional arousal [159]. An EDA signal can be decomposed into two different and non-redundant components: a phasic and tonic component [160]. The phasic component is the decomposition of the rapid movements of the signal, known as the skin conductance response (SCR), which commonly provides the features used in EDA-based studies to provide valuable information for many scientific research fields [161]. Special attention has been given to the approach by psychology and health-related studies [162]. In clinical analysis, SCR is used to assess pain, stress, schizophrenia, and peripheral neuropathy [159, 160]. In neuroscience and psychology, it is used to assess the subject's arousal levels [163]. For example, [164] used EDA signals to assess the stress of subjects in emulated real-life job scenarios, and [165] studied EDA to distinguish between stressful and calm conditions. [166] also analyzed the stress levels of a subject using the signal. Elsewhere, studies related to mental illness have utilized EDA signals, with [167] finding statistical evidence concerning the relationship between healthy patients and patients with bipolar disorder using features of EDA signals and [168] discovering significant correlations between EDA signals and engagement in dementia patients.

Most previous research has collected EDA signals in laboratory environments [169], where subjects are usually seated and often cautioned not to move the hand to which the electrodes are attached. However, recent applications have recorded EDA in environments where the users can walk freely and move their hands, such as daily-life settings and virtual reality (VR) environments. Notably, many wearable devices have been developed to enable the possibility of acquiring EDA signals in a daily-life scenario, leading [170] to propose a wearable EDA sensor for detecting drowsiness in drivers, [171] to analyze the affective state of children in everyday situations when interacting with robots, and Kim and Fesenmaier [172] to measure traveler emotions in real time during a four-day visit. Meanwhile, VR has been used to simulate environments where subjects can freely move and interact, which creates the sensation of being in the real world [173]. VR can display different scenarios to evoke emotions or provoke cognitive processes in the subject [77, 129] and has been used in case studies of, for example, social adaptation in social phobia contexts, the reduction of anxiety and pain, rehabilitation, and neurological diagnosis [174–178]. EDA has been used in VR experiments to examine sudomotor activity and arousal levels to assess anxiety and stress [179], conduct emotional assessments [180], and diagnose autism [181].

However, among the most significant issues concerning the use of EDA signals in daily-life and VR environments is the subject's movement during data collection. Although these technologies can offer an accurate environment for recording subject responses, the absence of control over the environments can impact EDA records. Most movements can cause interferences in the contact between the skin and the recording electrodes, producing major artifacts in EDA recordings [182]. [169] suggested that artifacts in EDA signals may conceal the existence of important correlations between the signal and the subject's arousal levels due to their heavy influence on the phasic component. Therefore, ensuring the quality of the signal in uncontrolled environments represents a critical challenge.

Most EDA-based experiments manually remove major artifacts using a human expert, because there is no robust and established methodology for automatically recognizing and correcting EDA signals. Artifacts can be manually corrected using various software, including Ledalab (www.ledalab.de) and SCRalyze [183]. However, manual correction has several disadvantages. First, it is a time-consuming and tedious task. Second, manual correction can introduce subjective human bias, with different experts correcting different signals. However, most critically, it cannot be applied in real-time or for short time periods without human intervention, as there is a demand for intelligent wearable devices that need to integrate a fully automated pipeline into the sensors. Examples of such systems include automatic systems for disorder recognition [181] adaptive therapies [178], mental health monitoring systems at home [165], driver drowsiness detection [170], and aesthetic evaluations [129].

Therefore, algorithms that can quickly detect and correct artifacts, ensuring data quality, appear essential for future applications of intelligent EDA-recording devices. However, works that develop automatic methods for removing artifacts remain limited [169, 182, 184–186] and present several limitations (see Section 2 for further details): i) The works that recognize artifacts detect whether a segment of a signal did or did not contain an artifact, but did not provide a continuous clean signal, which is needed to compute the phasic component and assess arousal, ii) the works that corrected signals did not compare their results with signals manually cleaned by experts, the most common method for removing artifacts, iii) previous works did not assess the impact of the correction on the phasic component, which is related to the emotional arousal dimension and represents the most important feature in the state-of-the-art approach, and iv) the performances of the different methods are not comparable because there is no public data benchmark. That is, no extant research has considered the development of a model that removes major EDA artifacts to provide a clean signal that does not have differences in terms of the phasic component with the signal that was cleaned manually by an expert.

This work develops an automatic recognition and correction algorithm for EDA signals, thus providing an artifact-free corrected signal that can be used in uncontrolled environments where users can freely walk and move their hands. This involves exploring two novel approaches: a long short-term memory neural networks (LSTM) in combination with a 1D convolutional neural networks (CNN), and a 2D CNN for spectrogram analysis. We compare these approaches with two state-of-the-art methods. A total of 74.46 h of EDA signal recordings were collected in a VR environment in which the 43 participants had to perform different tasks that required hand and body movements. The signals were manually corrected by two experts, generating an artifact-free signal for use as a ground truth. The labels obtained from the manual correction procedure were used to train and test the artifact recognition models. Next, automatic correction was performed on the artifacts detected. Finally, to measure the quality of the automatic corrections, the phasic component was evaluated pairwise with the automatic correction, the manual correction, and the original raw signal using two different decomposition algorithms, namely, CDA and cvxEDA.

The rest of this paper is organized as follows. Section 2 introduces the related literature. Section 3 describes the dataset's construction and the proposed methods for recognizing and correcting the artifacts. Section 4 presents the experimental results and provides a performance analysis of the proposed model. Section 5 discusses the findings, and Section 6 concludes the research.

3.1.1 Realted work

Several studies have considered EDA artifact recognition. For example, the work of [187] explored the recognition of EDA artifacts using a model based on four rules derived from the minimum and maximum range of the EDA signal or its temporal variation. However, the research on automatic detection of artifacts on EDA signals employing ML methodologies remains limited. Adopting a sampling frequency of 8 Hz, [182] detected motion artifacts in 5 s EDA segments and extracted different features from the raw EDA signal, including statistical variables (e.g., the mean, the maximum and minimum values of the data, and wavelet coefficients) to distinguish between artifacts and non-artifacts. A dataset with a duration of 130 min is used. The method achieved 96% accuracy using a support vector machine (SVM) model. However, the proportion of artifacts was not reported, and it should be considered when interpreting the performance of the model. [188] employed the same methodologies and objectives but used a larger dataset than other experimentations, including a total of 107.56 h between 13 participants. The data collected were based on ambulatory EDA signals with a sampling frequency 32 Hz that was later resampled to 8 Hz. Validation revealed a 98% true positive rate (TPR). However, the approach followed had certain limitations, such as recognizing artifacts using 5 s segments, a lack of evaluation of artifacts in the whole signal, and not providing a final corrected signal. In addition, the final dataset has an artifact percentage of 48.96%, which differs from the initial unbalanced percentage of artifacts (17%). [185] adopted a different approach, studying the use of unsupervised learning to identify artifacts from the raw signal, achieving competitive results compared to supervised learning. In addition, [189] also analysed an unsupervised approach using synthetic data as groundtruth. [186] presented recently a model that recognize segments of 5 s affected by artifacts with 94.7% of accuracy based on a ML model feeded by a new set of hand-crafted features. They compared the method with the methodologies of [182] and [187], outperforming the previous results. They collected both clean and corrupted EDA signal from immobile and moving hands, respectively, and their differences were used to create the groundtruth. However, they did not perform a correction of the artifacts providing reconstructed signals, which are needed for intelligent device systems, and did not analyse the implication of the artifact recognition on the phasic component.

In contrast, several works have studied the automatic correction of EDA signals without directly recognizing the artifact. That is, these methods modify the whole signal without needing to identify the artifact. Most contributions arrive from the field of signal processing, which has proposed using low-pass filters or exponential smoothing for artifact correction such as [190]. However, these approaches can modify certain segments of an EDA trace, which affects genuine physiological responses, creating more artifacts [169]. Other studies have used Stationary wavelet transform models to automatically remove artifacts in EDA signals. For example, the work of [184] models wavelet coefficients using a Gaussian mixture distribution. Their model required estimating three parameters using the expectation-maximization algorithm. Elsewhere, [191] made a breakthrough by studying the automatic model cvxEDA, which linearly decomposed the EDA signal into tonic components, phasic components, and a Gaussian noise term that represents the signal's white noise. Therefore, this algorithm enabled the direct decomposition of the EDA signal into two main components while simultaneously removing the noise term. This model is based on Bayesian statistics and convex optimization. [191] showed that cvxEDA outperforms CDA in terms of finely discriminating arousal levels. Furthermore, its low computational cost and efficiency has led to its use in other experiments e.g. [192, 193]. Meanwhile, [169] proposed a wavelet-based transformation based on the Stationary wavelet transform that used a zero-mean Laplace distribution to model the wavelet coefficients and only required estimating a single parameter. More recently, [194] used a deep convolutional autoencoder for automatic signal correction, which more effectively demonstrated the signal-to-noise ratio than previous methods. According to that work, "the ideal scenario would be having an extra reference clean EDA signal which then can be matched with the reconstructed signal to evaluate whether the reconstructed signal accurately recovers the underlying SCRs in the EDA signals without any distortions." However, only five subjects and 39 segments of the work include a clean EDA signal for evaluation, and the validation focused on the signal-to-noise ratio. Therefore, none of these works analyzed the implication of the correction in the phasic component of the signal to recover the underlying SCRs.

Although the correction methods used in previous works produced improvements in signal-to-noise quality, none of those studies validated their findings by using an EDA signal manually corrected by an expert as a ground truth. Having the clean signal as a reference can critically improve automated approaches by enabling not only the quantification of the existing artifacts via a comparison of raw and clean signals but also the evaluation of the correction via a comparison between the automatically corrected signal and the manually cleaned signal. Furthermore, this approach can compute the underlying phasic component of the clean signal and evaluate how the automatic correction impacts this component, the most important and common feature used in such studies. As such, emulating the manual corrections performed by experts must be the ultimate goal of ML and DL models given that most studies use manual correction for artifact correction [161].

Meanwhile, no previous research has combined artifact identification followed by signal correction in the same pipeline. In addition, none has been found that presented a precise characterization of motion artifacts (e.g., total number, duration, and percentage of the signal affected), which is especially important to characterize the noise levels of the signal used in each study and understand the differences on the results between studies. This might be due to the need for a manually cleaned signal to quantify the artifacts, a reconstruction that no study has included. Finally, previous studies have not made their models available for use by the scientific community, which limits the ability to produce comparisons between models. Furthermore, there is no benchmark public data, which would enable the same test data to be used in comparisons of novel methods with the state-of-the-art. As such, there are limitations when comparing the performances reported as the performances is related to the type and number of artifacts and the methodology used.

The most important existing studies on EDA signal filtering and artifact recognition are summarized in Table 3.1. The table presents the objective of the study, the methods used, the subject task, the target type, and the total EDA time used.

Concerning the model used to recognize artifacts in EDA, no previous work has explored the use of deep learning (DL) algorithms. DL is being a really important tool in order to classify and recognize patterns in different types of signals and images [195,196]. This tool has been applied in recent years to other physiological signals, such as an electrocardiogram (ECG) or electroencephalogram (EEG). Models such as U-Net [197], ResNet [198] or recurrent neural networks (RNNs) have shown good performances and versatility in different types of health-related problems. For example, Kyathanahally et al. [199] used two DL models for ghost artifact correction in EEG spectrograms. The first model classified whether the spectrogram had a spurious echo in it or not. The second model conducted a regression over the spectrogram to correct the artifact. This work shows promising results for the detection of ghost artifacts in EEG. With regard to ECG signals, the work of Bento et. al. [200] utilized two different DL classification models to study the classification of atrial fibrillation. The work of Liu et al. [201] performed a segmentation over hippocampal images using a 3D DenseNet to classify if certain subject sould suffer Alzheimer's disease. Finally, RNN was used by Antczak [202] to automatically denoise ECG signals. Various models were tested, including re-trained

Table 3.1: Summary description of each study related to the automatic correction or recognition of artifacts methodology in EDA signal, compared with the characteristics of the proposed work.

Study	**Objective**	**Used methods**	**Subject's task**	**Target**	**Performance evaluation**	**Time sample**
cvxEDA [191]	Signal correction	Bayesian statistics and convex optimization	Static breath and simulation	Raw signal	Increasement of arousal recognition vs. raw signal and CDA	5.1 h (30 subj.)
Shukla et al. [169]	Signal correction	Wavelet transform	Driving	Raw signal	Increasement of arousal recognition vs raw signal	20.11 h (15 subj.)
Chen et al. [184]	Signal correction	Wavelet transform	Physical, cognitive and emotional tasks	Raw signal	Artifact power evaluation metric vs other filter methods	81.5 h (32 subj.)
Taylor et al. [182]	Artifact recognition	Support Vector Machine	Physical, cognitive and emotional tasks	Expert labelling of 5 s window	% of windows with artifact detected	2.17 h (32 subj.)
Zhang et al. [185]	Artifact recognition	$k-$Nearest Neighbour	Static task in laboratory and real-life walking	Expert labelling of 5 s window	% of windows with artifact detected	23 h (21 subj.)
Our proposal	Artifact recognition + signal correction	(1) Classical ML (2) Recurrent Neural Networks (3) U-Net	VR tasks without movement restrictions	Conitinuous signal cleaned by expert	% artifact detected in EDA signal + signal generated phasic component vs corrected signal	74.46 h (44 subj.)

ones, and results improved for the pre-trained models with artificial ECG signals.

In summary, previous research did not assess artifact recognition and correction in the same workflow. Furthermore, no work has studied the comparison of automatic correction algorithms against an expert-corrected signal. Therefore, previous studies do not provide a final cleaned signal that emulates the type of correction performed by an expert.

3.1.2 Objectives

In this work, we develop an automatic recognition and correction algorithm for EDA signals, thus providing an artifact-free corrected signal that can be used in uncontrolled environments where users can freely walk and move their hands. To this end, we explore two approaches from the DL field that have not previously been used for artifact recognition in EDA signals. In addition, we compare them with a state-of-the-art ML method. A total of 74.46 h of EDA signal recordings were collected from 44 participants in an VR environment where the participants had to perform different tasks that required hand and body movements. The signals were corrected by two experts, generating an artifact-free signal that was used as a groundtruth. The labels obtained from the manual correction procedure were used to train and test the artifact recognition models. Afterwards, an automatic correction was performed on the artifacts detected by the best recognition model. Finally, to measure the quality of the automatic correction, the phasic component was evaluated pairwise between the automatic correction, the manual correction, and the original raw signal, using two different decomposition algorithms, CDA and cvxEDA. Therefore, the main objective of this work is to automatically recognize and correct EDA artifacts, achieving an automatically corrected signal that is indistinguishable from expert manual correction in terms of phasic components. This model can be used in intelligent wearable devices for monitoring human cognitive-emotional states for healthcare services without human intervention.

3.2 Materials and Methods

3.2.1 Participants

A group of 43 volunteers (13 females and 30 males) was recruited to participate in the experiment. The mean age of the group was 37.52 (SD = 8.38). The following inclusion criteria were applied: age between 18 and 50 years, Spanish nationality, and no previous VR experience. Before the subject's participation, they received documentary information about the study and gave their informed consent for their involvement. All methods and experimental protocols were performed according to The Code of Ethics of the World Medical Association (declaration of Helsinki), and the experimental protocol was approved by the ethics committee of the Universitat Politècnica de València (P4_18_06_19).

3.2.2 Data collection: EDABE dataset

We collected and published the Electrodermal activity artifact correction benchmark (EDABE) dataset [203], which includes raw electrodermal activity signals and the signals reconstructed via manual correction for use as a ground truth. To the best of our knowledge, this is the first public dataset, enabling comparison of methods. The EDABE dataset includes a total of 74.46 h of EDA recording affected by motion artifacts from the 43 subjects. It is divided into a training set with 33 subjects (56.27 h) and a test set with 10 subjects (18.19 h). We propose the adoption of the area under the curve (AUC) metric for evaluation on the test set. Given the dataset includes unbalanced classes, the AUC metric provides a more robust measure for future comparisons utilizing this dataset.

The data were collected during a VR study that had the objective of inducing stress in the subject by simulating daily situations at work in a virtual environment. The participants had to perform different tasks in the virtual scenario to achieve this objective. First, subjects were placed in an office setting, where they talked to a virtual avatar about issues related to work and personal life. Then, the subjects were moved to another scenario, a meeting with five virtual avatars in which they had to actively participate in decision making.

In all the settings that required conversations with the avatars, the subjects were able to choose between the four options displayed on the lower part of the screen. Finally, the participants played three different minigames. The first minigame involved extinguishing a fire in a virtual forest as fast as possible. The second minigame entailed reorganizing a pipe to allow water to flow through it in the minimum possible time. In the last minigame, the subjects had to complete a maze while simultaneously solving simple arithmetic equations as a parallel task. The faster the subjects solved both problems, the higher their score. In all three minigames, the participants had to move both of their hands to complete the games. As such, the EDA signal became noisier in the minigames section due to the induced stress and the subjects' rapid movements.

The subjects performed the VR scenario with a HTC Vive Pro-eye head mounted display working at 90 Hz refresh rate with 1440×1600 pixels per eye and a field of view of 110°. EDA data were recorded at a sampling frequency of 128 Hz using a Shimmer3 together with the Consensys software. A total, 43 EDA signals were collected. The average experiment duration was 104 ± 8 min, producing a total of 74.46 h of signals. The virtual environment is developed in Unity3D platform.

3.2.3 Methodology overview

The proposed methodology is summarized in Fig. 3.1. First, two experts corrected the EDA signals to provide the ground truth. Next, two state-of-the-art and two new models fitted over the training set were developed: i) [182], ii) [186], iii) an LSTM with a 1D CNN, and iv) a 2D CNN that analyze the signal's spectrogram. Following training and validation, the models were evaluated using the test set, with different classification metrics evaluated over each test signal. The algorithm that achieved the highest Kappa and TPR was selected as the best model.

Second, a fully automatic signal correction pipeline was developed. Artifacts were identified among the EDA signals using the best model. Then, a regression model was used to correct the detected artifacts to provide a final clean signal. Finally, the phasic component was calculated using the CDA and cvxEDA algorithms.

Validation of the complete pipeline involved comparing the phasic component of the three signals, namely, the raw signal, the automatic correction, and the expert manual correction (i.e., the ground truth). The similarity between the three signals was analyzed over the results of different regression metrics applied to each signal, namely, root mean square error (RMSE), mean absolute error (MAE), and cross-correlation. An ANOVA with a post-hoc analysis evaluated the differences between the phasic component of the signals.

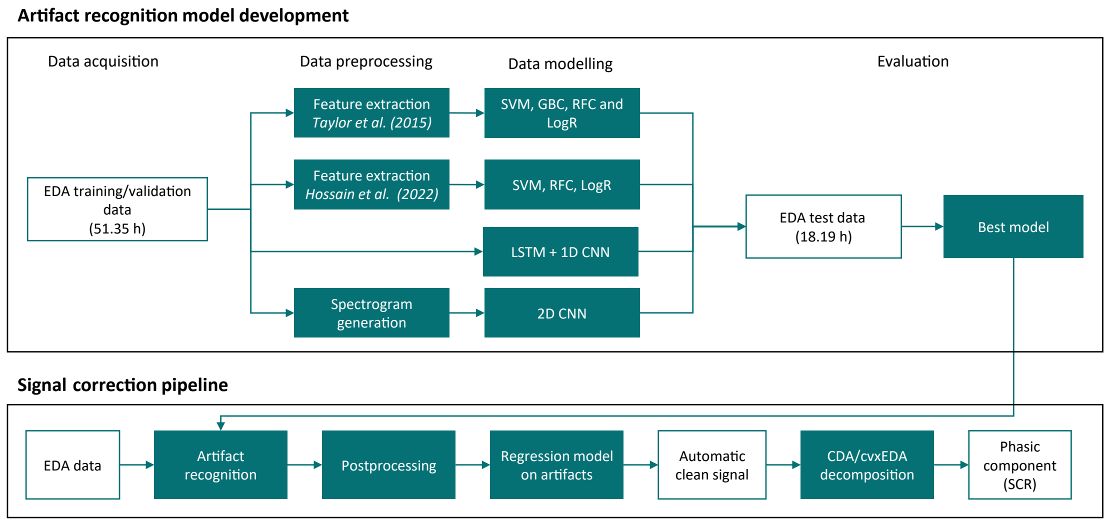


Figure 3.1: Schematic representing the artifact recognition and correction pipeline.

3.2.4 Expert artifact correction

The following procedure was used to obtain the manual correction of the signal. The expert cleaned the signal using Ledalab software, which allowed them to visualize the complete EDA signal and indicate, in the signal itself, in which sample the artifact started and ended. Ledalab allows the manual correction through different interpolations as linear or spline, allowing the expert to choose between the one that best suits the segment signal affected. The expert then performed an automatic interpolation on the signal, correcting the parts of it that were determined to be artifacts according to their own criteria. Ledalab recorded the corrected samples, thereby collecting the artifact samples. These data were subsequently used as labels to perform a binary classification that divided the samples into “artifact” and “non-artifact” samples.

One expert corrected 21 signals and the other corrected 22 signals, of which 33 were randomly assigned to the training set and 10 to the test set, representing 56.27 h and 18.19 h of EDA signal. The labels for each corrected signal were used to produce a descriptive-artifact analysis table.

3.2.5 Artifact recognition models

This work proposes four ML and DL classification algorithms. The first two methods replicates the methodology described by [182] and [186]. The four methodologies share the same target processing, assigning artifact or non-artifact label according with the percentage of artifacts in a 0.5 s segment. If more than 50% of the segment was labeled as an artifact, the sample of 0.5 s was labeled as an artifact; otherwise, it was labeled as non-artifact. All the models were fitted using the training set. As a filter, signals with an artifact percentage below 1% were removed, leaving 51.35 h of EDA signal to train the three models.

Upon training all four models, we conducted a test evaluation of the models that collected the mean values of different metrics, including accuracy, Kappa, TPR, and true negative ratio (TNR). Due to the considerable imbalance between the proportion of artifacts and non-artifacts, the Kappa score and TPR were selected to evaluate artifact detection performance. Once the best model was selected, we applied post-processing to the labeling provided by the model. This involved re-labeling the signal segment between two artifacts as an artifact if they were separated by less than a certain time threshold, with the aim of merging nearby artifacts. The time threshold used was fixed at 2 s. Subsequently, an additional metric was implemented, namely, the percentage of artifacts detected. This metric was used because artifacts are not single points but sets of samples with a time duration. As such, this metric measures the percentage of artifact detection. To consider a detection valid, we analyzed the percentage of the duration of the artifact that the model labels an artifact. If this percentage exceeded a threshold value, the corresponding detection was considered correct.

3.2.5.1 Taylor et al. model

The first method [182] is based on the extraction of several hand-crafted features from the raw EDA. The segments of 0.5 s are processed obtaining several types of features. The first is statistical features such as the minimum, maximum, mean, median, standard deviation and range. These statistical features were also computed over the first and second derivative of the segment. The same process is repeated for a low-pass filter of the signal with a frequency threshold of 16 Hz and to its first and second derivative. The last set of features was achieved from the computation of wavelet decomposition using Harr window of level three. From each level, the mean, median, maximum, standard deviation and number of coefficients above zero is computed. A total set of 62 features were obtained.

A backward feature selection (BFS) method based on SVC was used to select the best 40 features. Afterwards three different models were used, logistic regression (LogR), random forest classifier (RFC) and SVC. A parameter tunning was performed over each model to obtain the best hyperparameters, validating it through a group cross-validation of 5 folds. This type of cross-validation method was selected to ensure that the samples that belong to the same subject were not simultaneously present in train and validation split. The parameters used in the grid were 0.01, 0.1, 1, 10 and 100 for C in LogR; 200, 400 and 600 estimators, 10, 30 and 50 max. depth for RFC model; 1, 10, 100 and 1000 for C and 0.001, 0.01, 0.1, and 1 for Gamma in SVM model. The model with highest accuracy was selected as the best model.

3.2.5.2 Hossain et al. model

The second model reproduces [186] methodology. In our case, the model extracted the features and recognized wheter or not an artifact was present in a signal segment of 0.5 s instead of 5 s to produce comparable results. The computed features can be divided into three groups. First, statistical features such as the mean, median, standard deviation, minimum, maximum, range and shannon entropy from the raw signal and its first and second derivatives. These characteristics are also computed from the phasic component of the EDA signal. Second, autoregresive features were obtained from the coefficients of an autoregressive model over the 0.5 s signal segments, excluding the interception coefficient but adding the error variance. These type of features had also been used in other works related with time signal analysis [204, 205]. Finally, time-frequency features that were based in two different time-frequency transformations: variable frequency complex demodulation (VFCDM) [206] and wavelet. VFCDM was applied to the signal segment using four different frequencies: 64 Hz, 48 Hz, 32 Hz and 16 Hz. Standard deviation and mean were computed from this decomposition. From the wavelet decomposition, a three-level wavelet decomposition using Haar window is used. Mean, median, standard devation and range of each level is obtained for each level. A total of 50 characteristics were obtained.

Following the original work, a BFS based on RFC was used to select the best 40 features. The input data were processed using standard scaler and min-max normalization. Parameter tuning was implemented using group cross-validation of 5 folds. The studied models were SVM, gradient boosting classifier (GBC), RFC and LogR. The parameters used in the grid were 0.01, 0.1, 1, 10 and 100 for C in LogR; 200, 400 and 600 estimators, 0.01 and 0.1 learning rate and 3, 5 and 10 max depth for GBC; 200, 400 and 600 estimators, 10, 30 and 50 max. depth for RFC model; 1, 10, 100 and 1000 for C and 0.001, 0.01, 0.1, and 1 for Gamma in SVM model. Highest accuracy defined the best model.

3.2.5.3 LSTM-1D CNN

This section proposes a novel model that implemented artifact detection in the last 0.5 s of a 5 s signal segment. This model's main purpose is to learn from the signal's temporal evolution. The architecture of this model was inspired by the work of [200] and [202], who both used CNN and LSTM to extract features from a raw ECG signal. Our work uses a set of LSTM layers in combination with 1D CNN layers.

Fig. 3.2 details the model architecture. Its first two layers were LSTM layers of 16 neurons that returned the hidden state output for each input time step. Subsequently, the network included four convolutional levels, each of which featured three convolutional layers with a batch-normalization operation performed after each convolution. Finally, each level included a dropout value of 0.05 and a max-pooling operation of size 2. The numbers of filters in each level were 32, 64, 128, and 256; kernel size was 5. Finally, the model featured two fully connected layers of 256 and 16 neurons and a final fully connected layer comprising a single perceptron with a sigmoid activation function. The model was trained with the rmsprop optimizer at a learning rate of 5×10^{-5} and a batch size of 16. Due to the imbalance, the cost function used to train the model was the Dice-Sørensen coefficient (DSC). The model had an early stopping threshold of 30 epochs. The percentage of artifacts in the training set was 12.60%. No filter was applied to the raw signal. For each 5 s segment, min-max scaling was applied.

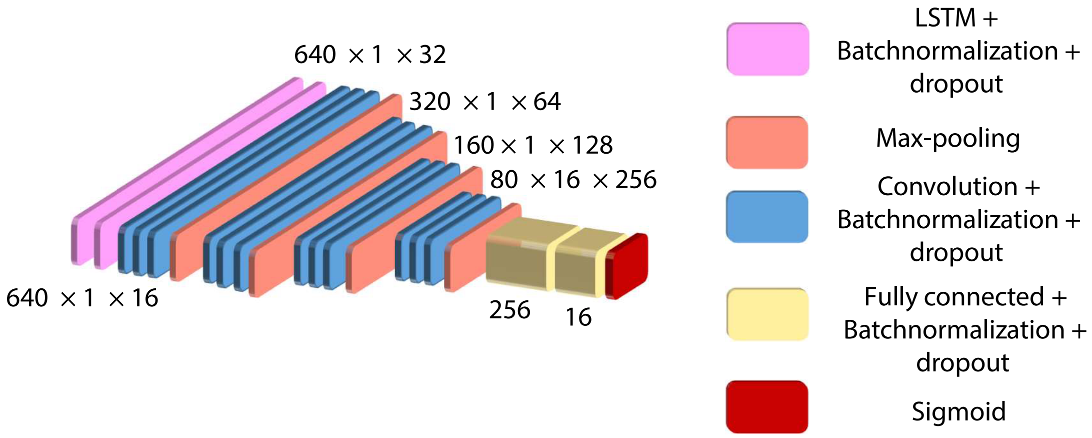

640 × 1 × 32
320 × 1 × 64
160 × 1 × 128
80 × 16 × 256
640 × 1 × 16
256
16
LSTM + Batchnormalization + dropout
Max-pooling
Convolution + Batchnormalization + dropout
Fully connected + Batchnormalization + dropout
Sigmoid

Figure 3.2: Schematic representation of the architecture used for raw signal classification and LSTM-1D CNN model.

3.2.5.4 Spectrogram and 2D CNN

The last proposed approach involved studying the recognition of artifacts via spectrogram artifact classification and segmentation. First, a spectrogram of each segment of 32 s of signal was created using fast fourier transform (FFT) with size 4096. Then, two consecutive models were used for the temporal segmentation of artifacts. The first model was an image classification model that classified a spectrogram as having an artifact or not. The second model was an image segmentation model that created a temporal segmentation inside the spectrogram to find the artifacts. This second model only studied the spectrograms classified as containing an artifact by the first spectrogram classification model. This model combination was based on the work of [199], and both models were based in 2D CNN layers. An overview of the pipeline appears in Fig. 3.3.

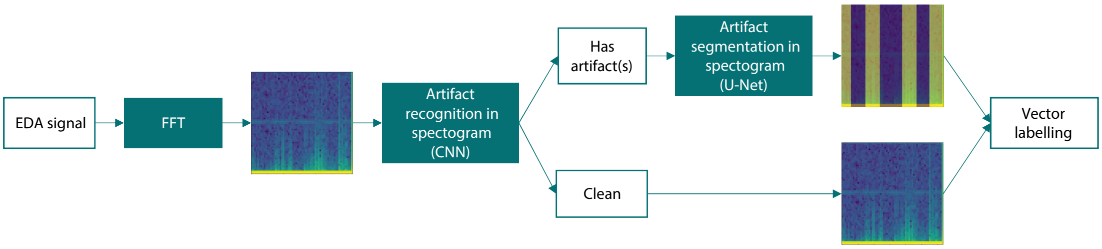

EDA signal
FFT
Artifact recognition in spectogram (CNN)
Has artifact(s)
Artifact segmentation in spectogram (U-Net)
Clean
Vector labelling

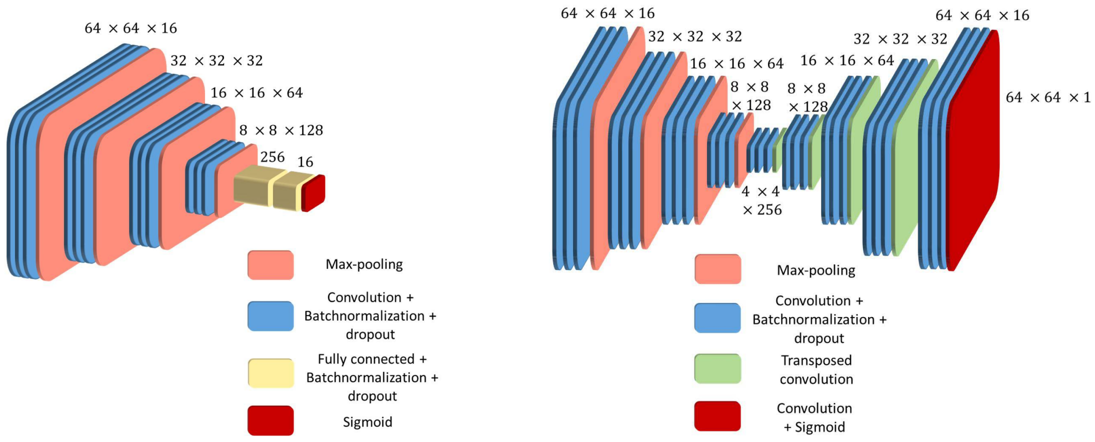
Figure 3.3: Scheme of the followed methodology for the detection and segmentation of EDA artifacts in the spectogram.

To obtain the spectrogram of a signal segment, the FFT algorithm was used. Using an FFT of size 4096, a resolution of 64 samples was achieved. To obtain the squared matrix, the time segments of each signal were divided into 32 s segments. A matrix representation with the dimensions 64 × 64 was obtained. In these representations, the vertical axis represents the frequencies in Hz, and the horizontal axis shows the temporal information in seconds. The spectrograms were obtained with a 50% overlap.

The classification model was a set of CNN layers used to perform image artifact recognition. The spectrogram was classified as containing an artifact if this percentage exceeded 0% based on a comparison with the ground truth. Otherwise, the spectrogram would be classified as clean. This binarization was used for labeling by the spectrogram classification model. The model architecture comprised four convolutional levels featuring between 16 and 128 filters, as Fig. 3.4(a) shows. The fully connected layers in the last two levels of the model had a dropout rate with a value of 0.5. The model's cost function was binary cross-entropy.

In contrast, the segmentation model followed a U-Net architecture, as Fig. 3.4(b) shows. The target image was a binary image in which the label 1 indicated an artifact. Therefore, the artifact was represented as a vertical segment in the spectrogram, with the width being the temporal segmentation of the artifact demonstrated by Fig 3.3. This pre-processing procedure produced the binary artifact mask image that was model's target. A maximum of 256 filters was used by the segmentation model. The kernel size for all CNNs was set to 5 × 5, and the dropout rate of the convolutional levels was set to 0.05. The model's cost function was calculated as the mean of DSC and binary cross-entropy. Using Adam optimizer with a batch size of 4, the learning rate for both models was 1×10^{-4}, and both models had an early stopping threshold of 30 epochs.

The total percentage of spectrograms with artifacts in the training set of the classification spectrogram model was 45.38%. Considering the spectrograms that contained an artifact, the total number of pixels identified as belonging to an artifact produced a total percentage of artifact pixels of 39.80%. The data introduced in the two models was a set of min-max normalized spectrograms with the dimensions 64×64. To increase the size of the training dataset and achieve a higher degree of model generalizability and robustness, the two models were trained using data augmentation technique [207]. For this, we implemented two different types of transformation. The first involved defining random vertical or horizontal lines equal to zero that hide—at random—certain pixels in the spectrogram. The minimum and maximum threshold numbers of hidden pixels were 256 and 1024. The second transformation was the translation of the spectrogram



(a) Spectogram artifact classification architecture

(b) Spectogram artifact segmentation architecture

Figure 3.4: Architecture of the two models included in the spectrograms and 2D CNNs. Image (a) shows the artifact classification model, and image (b) shows the model that achieved the segmentation of the artifacts in the spectogram.

image via a random vertical and horizontal pixel distance. The minimum and maximum threshold distances defined were 4 and 16 pixels. All the images in the dataset suffered both types of transformation, increasing the size of the dataset three times.

3.2.6 Artifact correction

Following the artifact recognition task, a regression model was developed to correct the detected artifacts via the samples of signals labeled artifacts. This automatic correction process combined two interpolation methods. The first was a linear interpolation between the beginning and the end of the artifact. The second involved obtaining a polynomial of degree 8. The first and last samples of the artifact were taken to obtain this polynomial, and six additional internal and evenly spaced samples were considered. The methods produced a set of points for each sample labeled an artifact. Finally, the techniques were averaged for each point of the artifact to combine the corrections performed using the linear and nonlinear approaches. This approach partially reproduced the methodology involving the use of the Ledalab software. The method used in this work combines the two approaches, with the linear fit capturing the tendency of the artifact segment and a $8th$ degree polynomial estimation to adjust the interpolation to the non-linearity of the EDA signal. Subsequently, a simple moving average of eight samples was implemented. The simple moving average was applied from 0.125 *s* before the beginning of the corrected artifact to 0.125 *s* after the end of the artifact to smoothen the joint between the corrected artifact segment and the original EDA signal.

A set of metrics was computed to evaluate the quality of the automatic correction. We analyzed differences in terms of the phasic component between (1) the raw signal, (2) the automatically corrected signal, and (3) the signal manually corrected by experts. We focused on phasic component because it assessed the sympathetic activity and the central meaning of EDA is revealed by its peaks [160]. To probe the robustness of the proposed methodology, we obtained the phasic component using two different approaches: the CDA (using the Ledapy library) and the cvxEDA algorithms. The metrics compared the three phasic signals by pairs, and the computed metrics were the RMSE, MAE, cross-correlation, and the difference in the area under the curve (DAUC). Furthermore, the phasic components of the signals were segmented into intervals of 5 *min*, upon which the mean could be computed. We analyzed the distribution of the means among the three signals using a one-way ANOVA test, performing a post-hoc analysis by pairs to observe statistical differences between them. The hypothesis considered is that if the automatic correction simulates the manual correction, no differences would be observed between them, while differences would be observed between the raw signal and the two corrections.

3.3 Results

3.3.1 Signal and artifact description

Table 3.2 shows the descriptive analysis of the artifacts identified considering the train and test stes, and the complete dataset. The mean artifact presence percentage was $10.63 \pm 11.59\%$.

Table 3.2: Descriptive features for the artifacts extracted from all signals. Metrics are shown as mean and standard deviation per participant. (*) Samples are computed considering a target each 0.5 *s*.

	Artifact duration (s)	**Number of artifacts**	**Signal affected (%)**	**First artifact (s)**	**Minimum artifact duration (s)**	**Time between artifacts (s)**	**Total samples with artifact***	**Total samples***
Train	5.37 ± 3.59	113.48 ± 97.12	9.97 ± 11.80	86.13 ± 173.98	1.08 ± 0.70	169.35 ± 291.63	44669	405194
Test	5.14 ± 3.01	182.30 ± 86.71	12.81 ± 10.57	45.65 ± 22.36	0.73 ± 0.55	48.47 ± 44.01	18246	130962
Complete dataset	5.22 ± 3.56	129.49 ± 99.16	10.63 ± 11.59	76.72 ± 153.75	0.88 ± 0.53	89.74 ± 125.76	62915	536156

3.3.2 Artifact recognition

Upon training and validating the four different approaches, the models were evaluated on the test set (18.19 h of recording), with the performance calculated via a binary classification each 0.5 s. Therefore, the models were tested on 130962 samples. The performance metrics shown in Table 3.3 are averaged across the test set, providing the mean and standard deviation for each metric.

Table 3.3: Evaluation of the different proposed approaches on the test set. Results appear as means and standard deviations. The model with the highest AUC, Kappa and TPR is highlighted in bold.

Model	**Accuracy**	**TPR**	**TNR**	**Kappa**	**AUC**	**DSC**
Taylor et al. [182]	0.91 ± 0.05	0.32 ± 0.13	0.98 ± 0.04	0.39 ± 0.09	0.65 ± 0.05	0.44 ± 0.12
Hossain et al. [186]	0.91 ± 0.05	0.38 ± 0.18	0.96 ± 0.08	0.42 ± 0.10	0.67 ± 0.06	0.47 ± 0.14
Raw signal and LSTM-1D CNN	0.88 ± 0.09	0.65 ± 0.16	0.89 ± 0.17	0.49 ± 0.08	0.76 ± 0.06	0.57 ± 0.07
Spectrogram and 2D CNN	0.87 ± 0.10	0.63 ± 0.17	0.87 ± 0.15	0.42 ± 0.09	0.75 ± 0.06	0.50 ± 0.11

Of the different ML models tested using the feature extraction and ML approach, the RFC was the best model following the features extracted from [182] whereas, the GBC outperformed the other models following the set of features of [186]. However, both performed worse than the DL approaches in terms of Kappa, TPR and AUC. The spectrogram and 2D CNN approach produced the second-best performances, achieving a TPR of 0.63 and a Kappa of 0.42. The best performance was achieved by the raw signal and LSTM-1D CNN approach, which achieved a TPR of 0.65 and a Kappa of 0.49. This performance is also corroborated by the AUC metric (0.76). This led to the selection of raw signal and LSTM-1D CNN approach as the model for recognizing artifacts to be implemented in the final pipeline.

The predictions of the raw signal and LSTM-1D CNN model were post-processed to render artifact recognition more accurate. This involved merging the artifacts separated by under 2 s. Table 3.4 shows an improvement in the mode's performance, producing a TPR of 0.72, a Kappa of 0.50 and an AUC of 0.79 in test set.

Next, we evaluated the percentage of artifacts detected in terms of different overlap thresholds. Fig. 3.5 shows a decrease in the percentage of detected artifacts according to the overlap ratio threshold. If we consider a 50% overlap threshold that is, considering identification as valid if the model classified the artifact at least half of the time the

Table 3.4: Evaluation of the raw signal and LSTM-1D CNN model predictions on the test set after artifact recognition post-processing. Results appear as means and standard deviations.

Model	Accuracy	TPR	TNR	Kappa	AUC	DSC
Raw signal and LSTM-1D CNN + post-processing	0.87 ± 0.10	0.72 ± 0.13	0.86 ± 0.18	0.50 ± 0.10	0.79 ± 0.06	0.58 ± 0.10

model detected 59.88% of the artifacts. In addition, if we considered a 20% threshold the model identification increased to 81.39%.

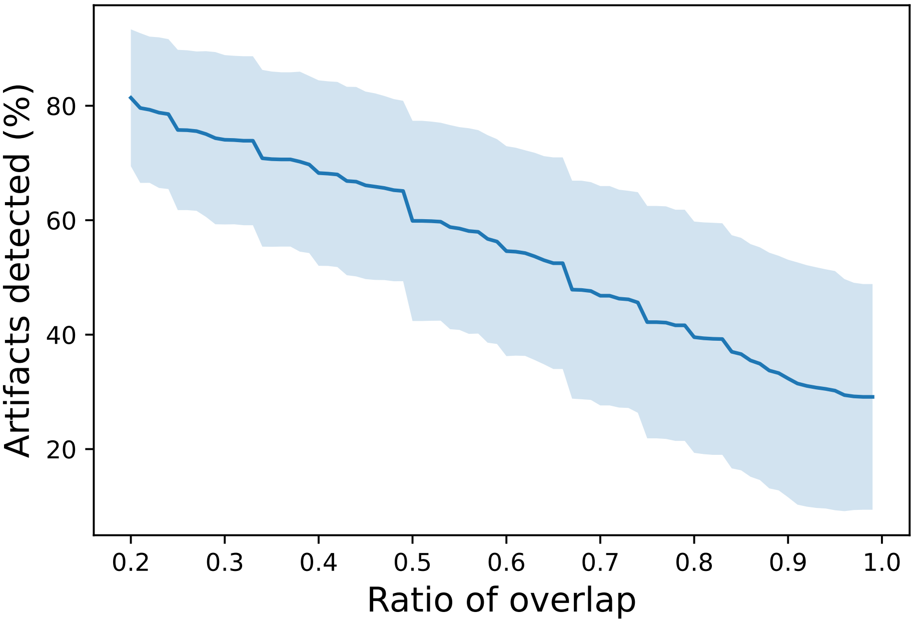


Figure 3.5: Evolution of percentage of artifacts detected in terms of the overlap ratio threshold. The line represents the average in the metric; its margin area, in light blue, indicates the standard deviation above and below the mean of the metric.

3.3.3 Artifact correction

Using the LSTM-1D CNN model with post-processing, a fully automated pipeline was implemented to the test signal data to obtain clean signals. This included a regression to interpolate the signal during the artifacts and a decomposition of the signal into phasic and tonic components. Fig. 3.6 shows the final interpolation result for a raw signal segment after the automatic correction process. The supplementary materials include the signals automatically corrected by the discussed algorithm.

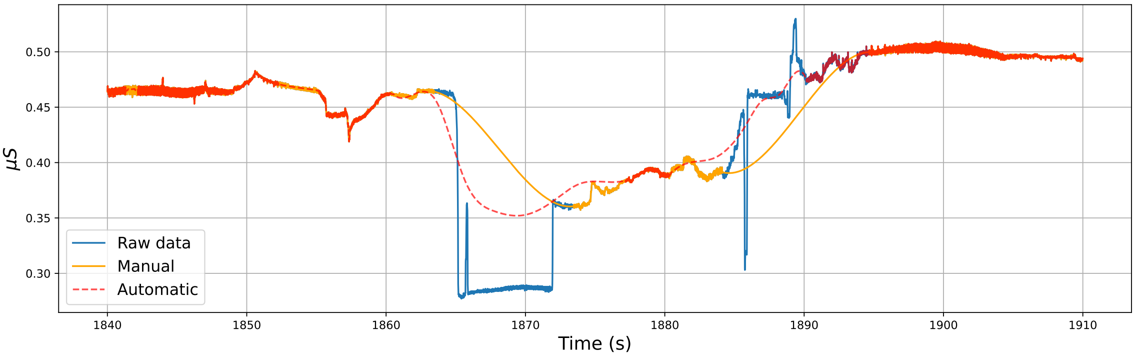


Figure 3.6: Automatic correction of a certain segment of an EDA signal. The blue line is the raw signal of the segment. The orange line is the manual correction performed by an expert, and the red line is the automatic correction performed by the artifact recognition and correction algorithm.

We validated the complete pipeline by comparing the phasic component of three signals: (1) the raw signal, (2) the automatic correction, and (3) the expert manual correction (as ground truth). This involved a pairwise evaluation of the signals. Table 3.5 shows that automatic and manual cleaning produced lower RMSE, MAE, and DAUC values according to both decomposition algorithms (CDA and cvxEDA). The ANOVA test did not find any statistical differences (p-value > 0.05) between automatic and manual corrections. In contrast, statistical differences (p-value < 0.05) were observed between the automatic cleaning and raw signal and between the manual cleaning and the raw signal. Fig. 3.7 shows boxplots of the values of the phasic components for each signal and decomposition analysis. In accordance with posthoc analysis, both signals demonstrate a higher similarity in the distribution of automatic and manual clean signals compared with the raw signals.

3.4 Discussion

This work aimed to develop a fully automatic pipeline for recognizing and correcting artifacts in EDA signals collected in uncontrolled scenarios involving hand and body movements. The work applied two new approaches using DL algorithms: an LSTM-1D

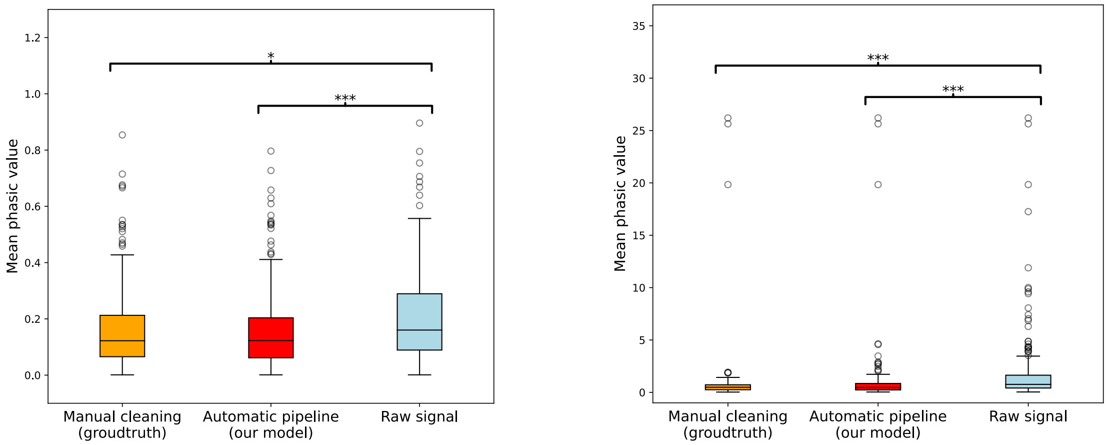
Table 3.5: Statistical metrics for the pairwise evaluation of the phasic components of the automatic corrections, the manually cleaned signals, and the raw signals. Results appear as means and standard deviations for each participant.

Algorithm	**Phasic component**	**RMSE**	**MAE**	**Cross correlation**	**DAUC**	p**-value**
CDA	Automatic and manual	0.146 ± 0.096	0.054 ± 0.033	0.772 ± 0.229	0.194 ± 0.184	0.427
	Automatic and raw signal	0.171 ± 0.108	0.068 ± 0.071	0.743 ± 0.216	0.246 ± 0.247	< 0.001 (***)
	Manual and raw signal	0.153 ± 0.102	0.064 ± 0.055	0.795 ± 0.186	0.377 ± 0.616	0.012 (*)
cvxEDA	Automatic and manual	0.339 ± 0.256	0.078 ± 0.039	0.633 ± 0.235	0.236 ± 0.168	0.246
	Automatic and raw signal	0.929 ± 0.786	0.272 ± 0.437	0.609 ± 0.207	0.478 ± 0.230	< 0.001 (***)
	Manual and raw signal	0.835 ± 0.809	0.255 ± 0.423	0.682 ± 0.278	0.317 ± 0.311	< 0.001 (***)

CNN applied to the raw signal and a 2D CNN applied to the spectrogram. The previous works of Taylor [182] and [186] were used as a benchmark.

This research contributes several novelties that build upon the state-of-the-art approaches. First, some previous research on artifact recognition [182, 185, 186] had detected whether a segments of a signal contained an artifact. However, they did not provide a final clean signal enabling computation of the phasic component. This could be critical because, for example, [186] analyzed segments of 5 s and, considering that many artifacts in our signals are shorter (see Table 3.2), this analysis could affect long segments of uncorrupted signal. Meanwhile, other studies had not recognized artifacts, instead aiming to directly correct signals using, for example, wavelet-based transformation [184] or convolutional autoencoders [194]. However, these works did not use manually reconstructed signals as a ground truth, which represents the objective of this study, that is, to emulate the reconstruction performed by an expert by providing an artifact-free signal.

This is the first work to develop a fully automatic pipeline with three steps: (1) artifact recognition each 0.5 s, (2) post-processing of artifact recognition, and (3) correction of the signal based on artifact identification. Meanwhile, by using as a ground truth a manual reconstruction of the signal, we have been able to assess the pipeline's



(a) Boxplot Ledapy

(b) Boxplot cvxEDA

Figure 3.7: Boxplot showing the distribution of different phasic values. Image (a) shows the comparison using the CDA decomposition method. Image (b) shows the results produced by the cvxEDA algorithm.

performance via a comparison of the automatic and manual corrections. Additionally, the dataset created contains 74.46 h of raw and manually reconstructed data, indicating labeling of more than 500.000 samples of 0.5 s. The data were collected from 43 different participants, ensuring the capacity to perform inter-subject extrapolations. The uncontrolled scenario used guaranteed the production of hand and body motion artifacts because participants needed to complete minigames causing major motion artifacts and simulating the real implementation conditions of intelligent EDA devices.

Notably, no previous work had analyzed the implications of automatic corrections for the phasic component of the signal, the most common feature used in studies because it relates to arousal [161]. This may be due to the need for the reconstruction of the signal to analyze the implications of the correction for the phasic component, information not contained in the majority of previous work. This work has analyzed the differences between the phasic component derived from our pipeline and the manual correction, demonstrating no differences between them. This novel result supports our pipeline as an emulation of human expert artifact correction.

Furthermore, this is the first artifact correction pipeline that is available for the use (and testing) of the scientific community and the first work that includes a dataset featuring raw data, manual reconstructions, and automated corrections. Thus, it represents a benchmark that can be used by future researchers to compare new methods and improvements using the same data, a current limitation on the state-of-the-art limitation that precludes comparison of the results because different data are used. These methodological improvements represent a breakthrough in the validation of recognition and correction algorithms for EDA signals.

We used two state-of-the-art methods as a benchmark. In the test set, both achieve the highest accuracy, but they present the lowest Kappa and AUC. It is because the TPR is relatively low, since [182] and [186] detects the 32% and 38% of the artifacts respectively. Notably, these results were worse than those presented by previous studies. There are two potential reasons for this discrepancy. First, the type of labeling used in the present work differs from that used in other studies. That is, other studies directly assigned a complete window of 5 s the label of artifact or not artifact, while we used the comparison with the manual correction to assign this label in segments of 0.5 s, which suppose an important increasement of the precision of the correction. Second, the imbalance of the current dataset (10.63%) exceeds that of previous experiments (e.g., 48.96% in the work of [188]). This could bias the performance and the results of previous works. Note that we use a dataset collected during a VR Serious Game, which is an actual use case of the pipeline, while as an example [186] create a specific protocol to generate the artifacts.

The two models using DL architectures outperformed the feature extraction and classical ML approach, achieving higher TPR, Kappa, AUC, and DSC values in the test set. Inspired by prior research on ECG denoising [200, 202], we investigated the use of a LSTM-1D CNN. Concurrently, the adoption of a 2D CNN was explored, drawing motivation from previous studies on Magnetic Resonance Spectroscopy denoising [199]. The best model was the raw signal and LSTM-1D CNN model, which achieved a final accuracy of 0.88, a Kappa value of 0.49, and a TPR value of 0.65. This represents a large increase in artifact detection performance relative to previous methodologies. Meanwhile, the spectrogram and 2D CNN model achieved a Kappa value of 0.42, a TPR value of 0.63, and total accuracy of 0.87. This model's performance was inferior to that of the raw signal and LSTM-1D CNN model, likely because 2D CNN was not optimized for the study of spectrogram images due to the non-local information that a spectrogram provides, with CNNs basing their knowledge on the local information contained in the data. More specialized models, such as spectral-CNN, could be implemented in future research to study the artifact detection problem in the EDA context.

To improve artifact recognition, a post-processing method was applied to the predictions of the raw signal and LSTM-1D CNN model. This post-processing improved artifact detection, as demonstrated the increased Kappa and TPR values (0.50 and 0.72, respectively). We also analyzed the percentage of artifacts included in the test set that were correctly identified by the model. Considering an identification valid if the model correctly labeled 50% of the artifact, the pipeline recognized 59.88% of the artifacts, with detection increasing to 81.39% with the use of a 20% threshold. Therefore, most artifacts were identified at least partially correctly, potentially because the model identifies the most aggressive segments of artifacts but not the entirety of the correction made by human experts.

Finally, the EDA signal was corrected using linear and polynomial regressions on the segments identified as artifacts. The automatic correction algorithm used in this work was designed to be similar to the type of manual correction enabled by Ledalab software. Although the results obtained fulfilled the initial objective, the type of automatic correction could be complemented or replaced by other correction methodologies. For example, the methodologies suggested by [184] or [169], who implemented wavelet transformation, lowpass filters [190], and the cvxEDA algorithm [191] could enrich the corrections made by the proposed algorithm.

The complete pipeline was evaluated based on the implications of the corrections for the phasic component. This involved using a one-way ANOVA with a post-hoc test to compare the three signals: (1) the raw signal, (2) the automatic correction, and (3) the manual correction by a human expert (ground truth). According to Table 3.5, there was no statistical difference (p-value > 0.05) between the phasic component produced by the automatic correction and the manual correction for either the CDA or cvxEDA algorithm. Furthermore, the type of correction performed was robust against the type of signal decomposition applied, showing similar results for the two algorithms. Meanwhile, statistically significant differences (p-value < 0.05) were observed between the phasic component of the raw signal and the manual correction, as well as between the raw signal and the automatic correction. This indicates that the automatic correction features less artifact noise than the raw signal (see Fig. 3.7). Other metrics, namely, RMSE, MAE, and DAUC, also showed that the phasic component of the automatic correction was closer to the phasic component of the manual correction than to the phasic component of the raw signal. Therefore, the results suggest that the automatic correction accurately simulates manual correction, independently of the decomposition algorithm used. These results support this paper's main objective of providing an artifact-free corrected signal that emulates manual correction by a human expert.

However, the study does have some limitations that must be addressed in future research. First, model results can be improved by including more experts for manual correction to reduce human bias. This would enrich the signal target and, therefore, the generalizability of the models. Second, the visual inspection and manual reconstruction can create an unrealistic morphology in the EDA signal, even if it is the standard practice in experiments. The manual cleaning aims to reduce the negative impact of the artifact on the signal and, in particular, on the phasic component, but it is not capable of reconstructing the real affected EDA. The alternative approach to obtain the artifact-free signal is to create a protocol that forces one hand to generate movements while the other is stationary, collecting data from both different locations simultaneously, as performed by [186]. Even if these protocols may have a low degree of ecological validity since it is an artificial task, and EDA signal can change depending on the location [208], the model must be tested considering this alternative groundtruth. Third, future research should evaluate the model in other types of environments and tasks because the specific movements performed can modify the form of the artifacts. Validating the methodology for EDA signals collected during other types of tasks would strengthen the model and demonstrate its applicability to other contexts as real-world experimentations. Moreover, the procedure has not been tested for signals from different EDA devices or those with frequencies below 128 Hz. The methods established here could be studied at different sampling frequencies to review their performance and generalizability. Finally, future experiments should consider researching the development of fine-tuned architectures for different models, which could improve their classification metrics. For example, generative-adversarial networks and reinforcement learning represent promising alternatives to the models demonstrated in this work.

3.5 Conclusion

We have developed a fully automatic pipeline for recognizing and correcting EDA motion artifacts, achieving a corrected signal that does not differ from manual correction by human experts in terms of phasic component. The recognition of the artifacts outperforms two previous state-of-the-art methods. These results show that EDA signal correction in scenarios that require body movements can be achieved automatically, findings that can enhance the use of EDA signals in future experiments conducted in uncontrolled environments, including immersive VR and real-world settings. These findings also provide encouragement for the development of intelligent devices for recognizing human emotional states for healthcare services without human intervention, including implementations in the contexts of disorder recognition, adaptive therapy, remote mental health monitoring systems, and driver drowsiness detection.

Data and model availability

The complete pipeline is available in
https://github.com/ASAPLableni/EDABE_LSTM_1DCNN. Furthermore, the EDABE dataset is publicly available in Mendeley Data for use as a benchmark in comparisons of the performance of future models and pipelines
https://data.mendeley.com/datasets/w8fxrg4pv5 [203].

Chapter 4

Developing conversational virtual humans for social emotion recognition based on large language models

Llanes-Jurado, J., Gómez-Zaragozá, L., Minissi, M. E., Alcañiz, M., & Marín-Morales, J. (2024). Developing conversational Virtual Humans for social emotion elicitation based on large language models. Expert Systems with Applications, 246, 123261. https://doi.org/10.1016/j.eswa.2024.123261.

Abstract

Emotions play a critical role in numerous processes, including, but not limited to, social interactions. Consequently, the ability to evoke and recognize emotions is a challenging task with widespread implications, notably in the field of mental health assessment systems. However, up until now, emotional elicitation methods have not utilized simulated open social conversations. Our study introduces a comprehensive Virtual Human (VH), equipped with a realistic avatar and conversational abilities based on a Large Language Model. This architecture integrates psychological constructs—such as personality, mood, and attitudes—with emotional facial expressions, lip synchronization, and voice synthesis. All these features are embedded into a modular, cognitively-inspired framework, specifically designed for voice-based semi-guided emotional conversations in real

time. The validation process involved an experiment with 64 participants interacting with six distinct VHs, each designed to provoke a different basic emotion. The system took an average of 4.44 seconds to generate the VH's response. Participants assessed the naturalness and realism of the conversation, scoring averages of 4.61 and 4.44 out of 7, respectively. The VHs successfully generated the intended emotional valence in the users, while arousal was not evoked, though it could be recognized in the VHs. Our findings underscore the feasibility of employing VHs within affective computing to elicit emotions in socially and ecologically valid contexts. This development holds significant potential for application in sectors such as health, education, and marketing, among others.

4.1 Introduction

Affective computing (AfC) explores the capacity for eliciting, recognizing, comprehending, and appropriately responding to human emotions. This interdisciplinary field merges insights from psychology, computer science, and biomedical engineering [209]. Emotions play a critical role in a variety of processes, including decision-making, creativity, and social interaction. Many works have focused on discerning an individual's emotional state through biomarkers, self-report questionnaires, and machine learning (ML) algorithms. Given that AfC probes directly into human behavior, its applications span across diverse sectors, such as marketing and education, with a special emphasis on healthcare. For instance, numerous studies have investigated depression recognition utilizing a variety of biosignals [210] and recently, there are different works that have used voice as a biomarker of the depression level of the subject [211, 212]. Notably, the data collection environment for these systems is of critical importance, as it must mimic real-life scenarios to trigger specific phenomena associated with depressive symptoms, which are linked with heightened negative emotional states [213]. Other research has focused on stress recognition by analyzing physiological responses, with an aim to enhance treatment strategies and potentially prevent these conditions [214]. Hence, the ability to evoke, comprehend, and recognize emotions poses a significant challenge, one that has far-reaching implications across numerous fields, particularly in enhancing human wellbeing.

Emotions can be quantitatively analyzed using Russell's circumplex model [4]. This model postulates that every emotion is a linear combination of two affective dimensions: arousal and valence. The arousal dimension delineates the individual psychophysiological activation related to the emotion, while the valence dimension quantifies the subjectively experienced positivity or negativity of the emotion [5]. This bifurcation results in four distinct regions of the model: high arousal and positive valence correlate with emotions such as happiness or excitement; high arousal and negative valence is indicative of angry; low arousal and negative valence is associated with sadness or depression; and finally, positive valence with low arousal is characteristic of a relaxed or contented state. Emotions induce measurable physiological and behavioral changes [215]. Research in voice patterns has risen in the past year in the subfield known as Speech Emotion Recognition, demonstrating a high capability for recognizing emotions [216], founding this conclusion in many different languages. Moreover, other biosignals has been used to model emotions such as eye-tracking, electrodermal activity (EDA), electrocardiogram (ECG), and electroencephalogram (EEG).

The majority of AfC research primarily centers on stimuli derived from image or audio sources. For instance, the International Affective Picture System (IAPS) dataset comprises a set of images depicting individuals, objects, or events, all of which have been standardized based on valence and arousal parameters [217]. Another example is the DEAP dataset, which encompasses a collection of 120 videos evaluated in terms of arousal and valence [218]. However, the pursuit of AfC research necessitates meticulous approximation to specific, realistic daily scenarios, thus underscoring the need to develop innovative ecological experimental tools. In the landscape of such advancements, virtual reality (VR) has emerged as a significant and promising tool. Currently, VR heralds a paradigm shift for behavioral research in psychological assessment. It allows a reproducible experience between different users with a high degree of presence, the sensation of 'being there'. Technological strides have facilitated the manifestation of VR across diverse platforms. One application is via powerwall screens or cave automatic virtual environment (CAVE) technologies. Although classified as semi-immersive VR, these platforms nonetheless deliver a high degree of immersion without isolating the physical body of the subjects [90, 131]. In the realm of immersive technology, head-mounted displays (HMDs) offer a distinct experience. These devices provide an unparalleled degree of immersion, isolating the user from the external world and simulating a complete virtual experience [132]. Owing to these characteristics, VR has emerged as a highly engaging and realistic medium for emotion elicitation. Numerous experiments employing this tool have successfully achieved their objectives in emotion elicitation. For instance, the study by [122] explored emotion elicitation through four different virtual environments, effectively validating their initial hypothesis by inducing varied emotional responses in alignment with each virtual environment. Similarly, the work of [219] demonstrated that VR could serve as a more efficient medium to elicit emotions such as anger or fear compared to desktop or non-immersive tools.

Another important attribute is that VR also allows the introduction of elements which are not possible to achieve in real world experimentation such as virtual humans (VH). VH are human-like characters which commonly interact through a computer screen and/or speakers. They exhibit human-like behaviours such as speech or gestures, but also other human characteristics as emotions, empathy or memory [220]. For the moment, a VH is a computer program which tries to simulate a human. There is a set of main elements that build a body and the mind of the VH. On the one hand, the body aims to provide real-time audiovisual content for the VH, enabling real-time interaction, as well as the senses to receive information from the users. On the other hand, the mind aims to provide the ability to interpret natural language, reasoning, creativity, and memory, as well as providing a life history, mood, motivations, attitudes, and content. In particular, one of the most challenging parts of a VH is the interaction with a subject through a verbal conversation. The VH not only has to generate coherent, meaningful and contextualized messages, it also has to remember messages or ideas of the conversation. The origin of this type of interactive tools is found in chatbots and later in conversational agents.

The first software that allows to interact with natural language were chatbots. They are an informatic system that can establish a conversation with one or more users through different communication channels as voice, text or visual language [99]. However, classical chatbots have a pre-defined sequence of answers to the different possible inputs. It is a bounded system which response is already settled. The use of the chatbots is very diverse. The first operational chatbot is found in *ELIZA* [100]. Since there, chatbots have been a very popular field of study. Many different algorithms have been used to improve the communication with a subject trying to overcome the past models. Some of them are *MegaHAL* [102] which is based in Markov's model basing its prediction in a probability distribution choosing between the most likely words for the answer. Chatbots finally evolved when artificial intelligent (AI) algorithms were incorporated to this field.

The use of intricate models capable of generating the most appropriate response to a given input has given rise to the concept of Conversational AI. Unlike chatbots, conversational AI algorithms are unbounded, with responses generated based on the input. These models undergo prior training with data from conversations, books, or other forms of text, enabling them to communicate with humans. Several strategies exist to augment the interaction with conversational AI and render it more human-like, primarily through the use of language models. These models have been employed to enhance the performance of chatbots by enabling them to generate more complex responses not merely extracted from a database, but based on them. Additionally, the work of [221] defines an avatar as intelligent if it demonstrates proactive actions, responsiveness to the environment, and social interaction with other users. Early language models applied to generate conversations emerged from the work of [222]. The model developed in this study could perform two distinct tasks; providing technical support for Ubuntu programming and generating Twitter conversations. Today, there are various conversational AI technologies in use, such as Alexa, Siri, or Cortana [223, 224]. GPT models by OpenAI [225] represents a significant advancement in this field, emerging as one of the most sophisticated language models to date. Given the type of training GPT underwent, it can introduce contexts that align with the model's response. This not only enables the delivery of generic yet precise messages, but it also allows for responses to be crafted according to certain guidelines. Consequently, it becomes possible to orchestrate a conversation wherein the AI can interact based on the context's instructions, such as generating extroverted or empathetic messages, thereby simulating personality traits. These AI models are nearing the ability to emulate a conversation in a typical social situation with another human [220]. For instance, the study by [226] validates the use of GPT-3 and Blender as a question-answer teacher through written dialogues. However, despite these advancements, there is still a lack of research on emotion elicitation and recognition using conversational AI.

Furthermore, a significant portion of studies involving the use of conversational AI have traditionally relied on interaction with the subject through written text displayed on a computer screen. This method of communication is somewhat unrealistic as it fails to replicate a real-life conversation between human beings. Currently, various AI models, such as voice synthesizers or audio transcription tools, could foster a more authentic conversation. Audio transcription models, which transcribe spoken language into written text, have been a focal point of recent research. Most of these models convert the audio input into a Mel spectrogram to capture non-trivial features necessary for transcription. This task encounters several challenges, including the punctuation of sentences with periods or question marks, and transcribing different languages within the same sentence. However, novel automatic speech recognition (ASR) models based on transformers such as Whisper are outperforming the task of transcribing voice [227]. Conversely, voice synthesizer models have seen more extensive research over the years, yet certain issues remain unresolved, such as achieving naturalness in the synthetic voice. For a realistic conversation, it is crucial that the model's voice closely imitates a human voice. Additionally, appropriate intonation for questions or for conveying sadness or happiness is essential. The voice should also respect certain pauses, as indicated by periods or commas. This challenge has recently been addressed by platforms like AWS Polly, which incorporates commands in the plain text to indicate pauses or to instruct the model to speak louder or faster. The integration of VR and the aforementioned AI models could potentially construct a VH. Currently, most research in this area is conducted in 2D, using screens or smartphones [228]. However, these mediums may not fully realize the potential degree of realism achievable with VR. The study by [229] investigated the difference between a chatbot displayed through a screen interface and the same chatbot displayed through a terminal interface. The results revealed that an improperly displayed avatar could compromise the naturalness of the communication. Similarly, the work of [230] compared face-to-face human interviews with face-to-VH interviews in healthcare decision-making. This study yielded promising results, with VH interviews achieving ratings comparable to those of human interviewers. Interestingly, VH even outperformed human interviewers in terms of confidence ratings. Therefore, elements such as voice or appearance significantly influence the naturalness of the user's experience. VR could potentially enhance the sense of realism in a conversation with a VH and also allows interaction with and control of the VH within the virtual environment.

Nonetheless, achieving a realistic user experience with a VH presents a set of challenges that are currently difficult to overcome. The primary obstacles involve replicating the physical behaviors and movements of a human body in the avatar, in a bid to make the VH as realistic as possible. One of the significant challenges is lip synchronization, a task that only a handful of AI models are currently able to handle. Hand movements and gestures, which are vital for conveying naturalness, pose a similar hurdle. Another vital aspect is the ability of the avatar to evoke and express different emotions. This feature is not only critical from an emotion recognition standpoint, but it also necessitates a responsive avatar that can alter its facial expressions and gestures in response to the subject's messages. The study by [231] reveals that while artificial faces have different effects compared to real faces, both can stimulate a neurological response in the subject. Furthermore, [232] demonstrates that realistic avatar faces can aid users in accurately identifying emotions more effectively than without them. Several solutions exist for off-streaming scenarios, but very few are available for real-time or live cases.

In the realm of VR, there are platforms that enable the deployment of an avatar capable of synchronizing lip movements with an audio file. One of the most sophisticated and realistic platforms is Nvidia's Omniverse. It features a library called Audio2Face, which facilitates lip synchronization. Unity is also in the process of developing the Salsa library for coordinating lips and gestures. However, at present, no complete program or library can combine all the required elements to replicate the physical behavior of a human through the avatar of a VH in real time.

To the best of our knowledge, no previous work has developed a comprehensive VH incorporating a realistic avatar, conversational abilities based on Language Models, embedded psychological constructs such as personality, mood or attitudes, along with emotional facial expressions, lip synchronization, and voice synthesis, all intended for use in real-time semi-guided conversations. The primary objective of this research is to develop and validate a set of VHs that combine all these technologies and demonstrate the ability to elicit emotions in humans. To this end, we utilize various technologies from companies including AWS, Google, OpenAI, and Nvidia. Upon constructing the VH, all the different steps involved in the conversation are validated through an experimental setup involving 64 participants. This paper presents an analysis of the real-time performance of the conversational pipeline, an evaluation of human-machine interactions, the naturalness and realism of the generated conversation, as well as the emotions elicited in subjects, and the ability to identify VH's emotion. Our results reinforce the viability of utilizing VHs in the field of affective computing to induce emotions in ecologically valid contexts. Such advancements provide a range of applications in sectors such as health, education, and marketing, among others.

This paper is structured as follows: Section 2 elucidates the experimental design and the tools utilized therein, along with the analytical methods employed. Section 3 presents the results derived from the aforementioned analysis. Section 4 discusses the work accomplished and the objectives met, and provides a discussion of the results. Lastly, Section 5 concludes the study and highlights the key findings.

4.2 Materials and methods

4.2.1 Participants

A group of 64 subjects was recruited to participate in the experiment. The mean age of the subjects was 31.956, SD = 10.339 years, including 31 males, 32 females and 1 non-binary gender. The group was established based on specified inclusion parameters: a) ages between 18 and 55 years, b) right-hand dominance (due to possible influence on EEG patterns [233]), c) normal vision, and d) Spanish as the first language, articulated with a Spanish accent. Exclusion conditions involved: a) pregnancy or nursing, b) presence of psychiatric disorders, and c) usage of psychotropic drugs. To ensure the subjects constituted a homogeneous group and they were in a healthy mental state, they were filtered by the Patient Health Questionnaire (PHQ-9). PHQ-9 is a standard psychometric test used to quantify levels of depression. Significant levels of depression would have affected the emotional responses. Only participants with a score lower or equal than 9 were included in the study.

Prior to their participation, they received documentary information on the study and gave their written consent for their involvement. The responses were anonymized and randomized to ensure the privacy of the information. The study obtained the ethical approval of the Ethical Committee of the Polytechnic University of Valencia (P06_25_07_2022).

4.2.2 Instrumentation

The experimentation room was equipped with a 6.36 × 2 meters white screen, which was used as a method of projection for semi-immersive, life-size VR. This screen was outfitted with two Optoma projectors (ZH400UST DLP), each with a resolution of 1920 × 1080 pixels, which after blending, provided a projection of 3600 × 1080 pixels. The room was also fitted with a Logitech 5.1 sound system, model z5500. The PC used has the following spec: CPU, Intel(R) Xeon(R) W-2265 CPU @ 3.50GHz (24), RAM, 131585MB, and GPU, Quadro RTX 6000/PCIe/SSE2.

During the experimental procedure, we collected a variety of data from each participant, encompassing speech, ET, and physiological measures such as EEG, ECG, and EDA. Speech data was gathered using a SYNCO G1 A1 wireless microphone system, with the recording device positioned on the participant's forehead. ET information was recorded using the Pupil Invisible glasses from PupilLabs. The glasses feature a frontal camera with a resolution of 1088 × 1080 pixels, operating at a frequency of 30 Hz. The device recorded a the subject's gaze, and additionally provided an accelerometer and gyroscope, all functioning at 200 Hz. EEG and ECG signals were simultaneously captured utilizing the B-Alert x10 system (Advanced Brain Monitoring, Inc., USA), recording at a frequency of 256 Hz. EEG sensors were strategically placed in frontal (Fz, F3, F4), central (Cz, C3, C4), and parietal (POz, P3, P4) regions, adhering to the international

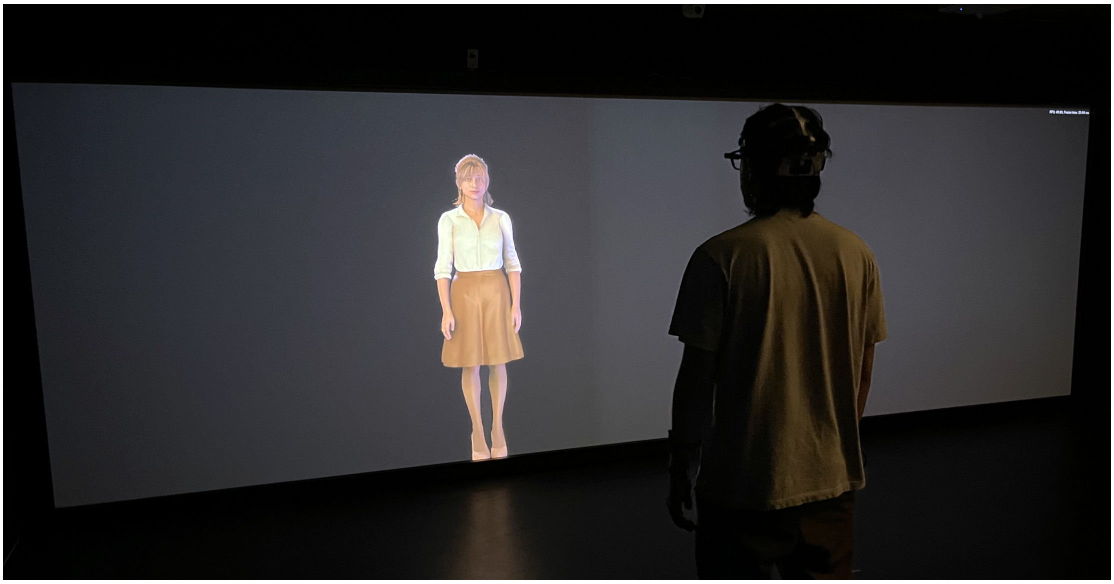


Figure 4.1: Photo of the experimentation room

10-20 electrode placement system on the participants' scalps. A pair of reference electrodes was positioned below the mastoid. Electrode conductivity was confirmed with an aim to maintain electrode impedance below 20 $k\Omega$. The ECG left lead was situated on the lowermost rib, while the right lead was placed on the right collarbone. Lastly, EDA signal was acquired using the Shimmer3 device, with a recording frequency of 128 Hz.

4.2.3 VH: General scheme

In our development of the VH, we employed a cognitive-inspired architectural design that divided the system into body and mind components. Drawing inspiration from the framework proposed by [220], the body component is further stratified into two subsystems encompassing senses and actions. A comprehensive overview of this design scheme is graphically represented in Fig 4.2. The VH's body action subsystem shows an avatar that is visually rendered with facial expressions tailored to reflect its emotional state. As participants vocalize, their audio is relayed in real-time to the VH's senses subsystem. Upon the termination of speech, a real-time silence detector module identifies the cessation and proceeds to route the audio to a transcription module, converting speech into text. This transcribed text is then transferred to the VH's mind system. Here, a conversation module leverages a LLM to generate responses. These responses take into consideration various factors including the VH's life-history, the contextual situation, attitudinal disposition, prevailing mood, and motivations. Additionally, it references the conversation memory to ensure temporal coherence. The crafted response text is then passed to the action subsystem of the VH's body. This response is synthesized into an audio file, broadcast through speakers, while the VH avatar's lip movements are synchronized to mimic authentic speech. Concurrently, the senses subsystem is primed to anticipate the forthcoming participant response, marking the commencement of the next iteration of the process. During the experiment, the behavioral and physiological responses of the subject are collected to be modelled in future steps of the work.

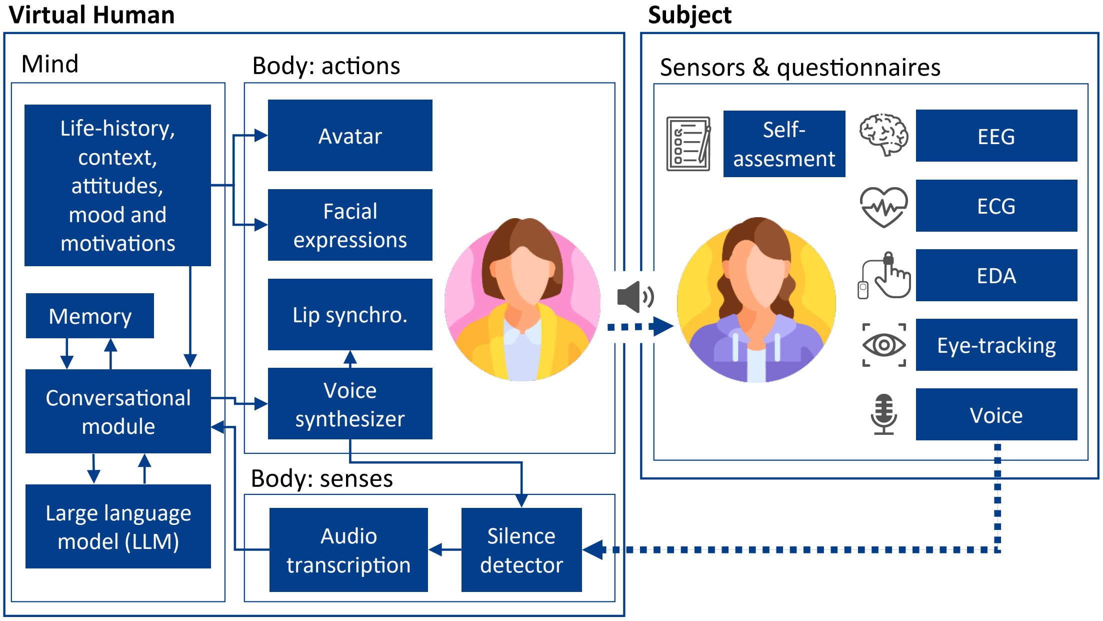


Figure 4.2: Scheme of the VH's architecture and the human-machine interaction

4.2.4 VH Body: Actions

The objective of the Actions subsystem, which is a part of the body component, is to produce real-time audiovisual content for the VH, enabling real-time conversational interaction.

4.2.4.1 Avatar

Four realistic avatars were acquired, comprising one male and one female, each with two variations: one in a casual outfit and the other in a semi-formal outfit. Five alternative emotional expressions (neutral, happy, sad, relaxed, and angry) were developed for each avatar, yielding a total of 20 avatars. The size of the avatars were designed according with the height of the population in Spain, 1.76 m and 1.62 m for men and women respectively. After acquiring the avatars in FBX format, they were imported into Autodesk Maya 2022 for rigging and the generation of blendshapes. Blendshapes, which are 3D mesh deformers, totaled 122 for the female avatar and 125 for the male avatar. Using these blendshapes, the polygonal mesh of the avatar could be manipulated to generate specific expressions such as modifying the corners of the mouth or the furrowing of eyebrows for a sad expression. Blendshapes can be grouped based on prominent and common facial parts such as eyebrows, eyes, mouth, ears, nose, cheeks, and neck, all of which have a considerable number of blendshapes for necessary modifications. Eye blinking was also animated using blendshapes. Next, the avatar's body geometry was cut at the base of the neck, thus dividing the avatar's geometry into two parts: the head and the body. Following this modification, an idle animation was created for the avatar and, once complete, it was exported in USD format using an Omniverse connector plugin for Autodesk Maya (Legacy). We adopted the definition of 'avatar' as a visual representation that includes technical aspects such as geometry, blendshapes, textures, materials, and rigging. This avatar can be controlled either by a physical human or by a VH, as described in [220]. It's worth noting that some authors consider an avatar to be only those representations controlled by a physical human and may regard it as a component of a larger agent [234].

4.2.4.2 Facial Expressions and Body Movement

The avatar's mouth, eyes, eyebrows, nose, and body movements were modulated based on the emotion to be elicited, utilizing the facial action coding system (FACS) [235]. The neutral avatar was designed with neutral features. The happy avatar was characterized by a moderate smile, a slight curved elevation of the eyebrows, open eyes, and a slight body sway. The relaxed avatar displayed a slight smile, a slight curved elevation of the eyebrows, a dilated nose, and contracted eyes. The angry avatar featured a very slight inverted smile, straight and lowered eyebrows, a contracted nose, and a furrowed brow accompanied by a body sway. The sad avatar exhibited a slight inverted smile, a raised brow, and contracted cheeks. All avatars presented a breathing animation. These facial expressions were designed based on the FACS and two online pre-tests were conducted with images of the faces on 102 subjects to refine the faces (further information in the supplementary materials). Table 4.1 show a summary of the traits for each mood.

Table 4.1: Facial and body movements design per emotion.

Emotion	**Arousal**	**Valence**	**Body Movements**	**Mouth**	**Eyebrows**	**Nose**	**Extra**
Neutral	Neutral	Neutral	Breathing	Neutral	Neutral	Neutral	-
Happy	High	Positive	Swaying	Moderate Smile	Slightly Raised Curved	Neutral	Open Eyes
Relaxed	Low	Positive	Breathing	Very Slight Smile	Slightly Raised Curved	Extended	Contracted Eyes
Angry	High	Negative	Swaying	Very Slight Inverted Smile	Lowered Straight	Contracted	Furrowed Brow
Sad	Low	Negative	Breathing	Slight Inverted Smile	Raised Brow	Neutral	Contracted Cheeks

4.2.4.3 Lip Synchronization

Nvidia Omniverse's Audio2Face app was utilized for real-time lip synchronization. The previously created USD file with the idle animation was imported and the ”Character Transfer” tool was used to animate the avatar's face. The Player Streaming feature was used to receive the synthesized VH voice. Texture modification could be achieved through the shader associated with each object's material, where the base color and opacity could be changed. To illuminate the scene, three rectangular lights were added: a rim light behind the character, tilted upwards with low intensity and a neutral color; another light at 45 degrees to the right of the character, tilted downward with high intensity and a warm color; and a third light at 45 degrees to the left of the character, tilted upwards with low intensity and a cold color. This lighting setup added realism to the character compared to the default lighting. The avatar was displayed directly in the full-screen mode of the application, prepared to synchronize the lips as the synthesized voice was played.

4.2.4.4 Voice synthesizer

The synthetic voice for the VH message was produced using Amazon Polly API, a service that leverages artificial intelligence models to convert text into an audio file in MP3 format. To provide a more natural and expressive prosody, we employed the speech synthesis markup language (SSML) for text input. This allowed us to introduce precise pauses at punctuation marks, such as periods and commas, with respective durations of 0.6 and 0.25 seconds. Two distinct voice identities were utilized: 'Lucia', a neural voice, provided the female articulation, while 'Enrique', a standard voice, was used for male avatars. Note that, at the time of the experiment, AWS did not offer a neural voice for Spanish males. These selections were made based on their representative qualities, contributing to a more immersive interaction with the VH.

4.2.5 VH Body: Senses

The senses subsystem of the VH is designed to endow it with a capacity to perceive its environment, focusing specifically on auditory perception.

4.2.5.1 Silence Detection

To facilitate real-time interaction with participants, we implemented an algorithm that detects the cessation of voice activity, signifying the end of a participant's speech. We utilized the silence detection model from the pyannote library [236], which ceases recording upon identifying a silence period exceeding one second. The onset and offset activation thresholds were both set at 0.5 s. The model commences silence checking three seconds after the onset of audio streaming. Should it identify an absence of voice activity lasting over a second, it ceases the streaming. Due to its operation in the background, this process does not cause any delay or interference with the participant's audio. The recorded audios are systematically stored in a designated folder for each participant and conversation.

4.2.5.2 Audio Transcription

Participant speech is transcribed into text, in the speaker's native language, utilizing Google Cloud's Speech-to-Text service. This service not only transcribes spoken words but also includes punctuation marks such as periods and question marks, enhancing the clarity and interpretability of the transcribed message for subsequent analysis. The model accepts audio input in WAV format and provides an output featuring multiple components of the recorded audio. If it identifies distinct segments of speech as separate messages, it will return each message individually. A post-processing function was employed to consolidate all individual messages generated by the AI model, thus preventing any overlap of messages.

4.2.6 VH Mind

The Mind system is designed to endow the VH with capabilities for natural language comprehension, rational thought, creativity, and memory. In addition, it furnishes the VH with a life history, context, attitudes, motivations, and mood.

4.2.6.1 Life-history, context, attitudes, mood and motivations

This module aims to create a narrative that provides a set of psychological features and context to the VH. In particular, it provides a life history (name, age, birthplace, current city, and profession), conversation's spatial context, motivation for the specific mood, and the VH attitude to engage in social interactions. The names used was selected from the list of most common names in Spain according to the Spanish National Statistics Institute, excluding compound names, trying to reduce any possible personal bias to a particular name. The current city was defined as Valencia (Spain) to provide spatial coherence, due to the experiment was performed in this city. The avatars with a non-neutral emotional state also include a motivation to engage in conversations on sports or cinema.

The constructed narratives conform to the following template: *"The following is a conversation with [name]. [name] is a [X]-year-old [nationality] [gender]. He/she was born in [birthplace] and currently lives in [current city]. [name] is [profession]. At the moment, [name] is [context]. Today, [name] is [mood] because [motivations]. Due to his/her [mood], [name] is [attitude].".* Given that this module establishes the VH's gender and mood state, this information is employed to manually select an avatar that coheres with these traits. To evade repetitiveness in opening and concluding phrases, each avatar is imbued with a distinct greeting and farewell. We hypothesize that the pre-defined mood embedded in the VH is capable of triggering the VH's emotions during the conversation and can elicit emotions in the subjects. Table 4.2 furnishes a summary of the parameters defined in this module for each VH.

4.2.6.2 Conversational module

The Conversational Module plays a crucial role in crafting the prompt that is sent to the large language model LLM. It integrates a variety of inputs, namely the emotional narrative (encompassing life-history, context, attitudes, mood, and motivation), the most recent transcription of the subject's audio, and, when applicable, the log of the conversation from the memory module. Notably, all these inputs are in Spanish. To personalize the dialogue template, the subject's name is manually added at the onset of the experiment. Using this information, a structured and context-sensitive prompt is assembled following the defined sequence. Subsequently, this prompt is transmitted to the LLM for processing. Interestingly, the LLM occasionally generates not just the VH's response but also anticipates the subsequent interaction from the participant due to the

Table 4.2: Summary of the different conversational parameters of each VH based on the mood.

	Neutral 1	**Neutral 2**	**Angry**	**Happy**	**Sad**	**Relaxed**
Life-history	Ana/David, 37 yo, Madrid, Pharmacist	Laura/Alejandro, 27 yo, Valladolid, Waiter/waitress	Marta/Jorge, 31 yo, León, Theater actor	María/Javier, 30 yo, Valencia, Teacher	Sara/Daniel, 25 yo, Ciudad Real, Studying psychology and retail associate	Lucía/Pablo, 21 yo, Bilbao, Book translator
Context	Subway stop	Tramway stop	Bus	Bar	Sitting on a park bench	Walking on the promenade
Motivation	-	-	She/He has been denied a salary increase and prices have gone up.	She/He has obtained a position as a primary school teacher in Valencia.	The landlord just called her/him to tell her/him that he needs the apartment in a month's time.	He/She is on vacation and has just left a spa.
Attitude	Kind, friendly and likes to talk about sport. Likes to chat and ask what sports the people she/he talks to like.	Kind, friendly and likes to talk about cinema. Likes to chat and ask what films the people she/he talks to like.	Needs to let off steam. She/He is not interested in the rest. Wants to talk about her/his anger.	Kind, friendly. Likes to talk and ask questions.	She/He does not feel like talking. She/He does not ask many questions.	Kind, friendly. Eager to talk and ask questions.
Greetings	Hello, my name is __ what is your name ?	Hello, my name is __ what is your name ?	My name is __ what is your name ?	Nice to meet you, my name is __ What is your name ?	Hello, my name is __ and yours ?	Hello, my name is __ what is your name ?
Farewall	I'm sorry, I have to go. We'll talk some other time. See you later!	I'm sorry, I have to go. We'll talk some other time. See you later!	Well, I have to go. Goodbye!	I'm sorry, I have to go. It was nice meeting you and talking with you. See you later!	I'm sorry, I have to go... See you later!	I'm sorry, I have to go. Nice talking with you. See you later!

number of token to generate is fixed. Hence, the module performs a post-processing step to filter out a single coherent response from the VH. Additionally, this module makes adjustments for specific words or expressions that the synthesis process struggles to handle accurately, such as "jaja" or "(laughs)". Table 4.3 presents an illustrative example of the prompt crafted by the Conversational Module:

4.2.6.3 Large Language Model

The LLM employed in this study is GPT-3 by OpenAI, specifically utilizing the *text-davinci-002* variant via API. The model had different parameters such as temperature, presence penalty, frequency penalty and maximum number of tokens which were meticulously calibrated. Temperature determines the output's level of focus and determinism; higher values make the output less deterministic. This parameter ranges from 0 to 1, and it was set to 0.9 in this study. The presence penalty parameter controls the introduction of new tokens and enhances the model's reluctance to discuss novel topics as the parameter value increases. Ranging from -2 to 2, the presence penalty was set to 0.9. The frequency penalty penalizes new tokens based on their existing frequency in the text, reducing the likelihood of the model repeating the same line verbatim. This

Table 4.3: Illustration of the prompt used to create a VH's response

> ”The following is a conversation with Marta. Marta is a 31 year old Spanish woman. Marta was born in León and is living in Valencia. Marta works as a theater actress. Marta is on a bus and is talking to the person sitting next to her. Today Marta is angry because she has asked for a salary increase and has not been granted it. This year the prices of electricity, rent and food have gone up and with Marta's salary it is difficult to make ends meet. In this situation, Marta cannot spend money on activities she likes to do in her free time, such as going to the gym or to the movies, because she has to save money to be able to pay for all the expenses. Today Marta is angry and needs to let off steam. That's why Marta is not interested in getting to know the people she talks to and only wants to tell them why she is angry.
> Marta: My name is Marta. What is your name?
> (SUBJECT'S NAME): (...)
> (CONVERSATION LOG)
> Marta: ”

parameter also varies between -2 and 2, and it was set to 0. Lastly, the maximum number of tokens indicates the quantity of new tokens generated in the language model's response. It can range from 1 up to the model's token size limit, which was 4000 in this case. For this study, the maximum number of tokens was set to 150. These parameter settings were established through an exhaustive iterative process involving numerous trials during the pre-pilot phase.

4.2.6.4 Memory

The memory module serves a pivotal role in preserving the conversation's log. This stored log is subsequently utilized by the Conversational Module to uphold the short-term temporal coherence, thereby ensuring the continuity and flow of the conversation.

4.2.7 Experimental procedure

In the initial phase of the experiment, participants were asked to fill out a battery of psychological questionnaires encompassing demographic information, PHQ-9, State-Trait Anxiety Inventory Questionnaire (STAI), and several other physiological assessment questionnaires, which will be used for further analysis. STAI was used to assess the emotional baseline of each participant. Following this, participants were equipped with the sensors described and proceeded to engage in six distinct conversations with a set of six VHs. They were displayed at a realistic scale via a projector encompassing an entire screen in a dimly-lit room, where the screen served as the primary source of light. This configuration helped in concentrating participants' attention on their communication with the VHs. The commencement of each interaction took place at a distance of 2.60 meters from the screen, although adjustments for cognitive and social comfort were allowed. The research team oversaw and managed the experiment from a separate location to minimize any potential influence on participants' responses.

The experimental procedure is shown in Figure 4.3. The initial two interactions involved neutral-emotion VHs, with the order of gender representation counterbalanced. Subsequently, participants engaged with four emotionally expressive VHs expressing happiness, relaxation, sadness, and anger, with the order and gender of these interactions again counterbalanced. Prior to each interaction, participants were provided with a situational context (e.g., "You are at a tramway stop and a person initiates a conversation with you"). The VH initiated the dialogue with a greeting, enabling participants to partake in open-ended conversations absent of specific guidelines or objectives. Interactions could be concluded by participants bidding farewell to the VH, prompting the experimenter to manually terminate the task. Alternatively, after a maximum duration of 4 minutes, the VH automatically bid goodbye, and the task was concluded by the experimenter.

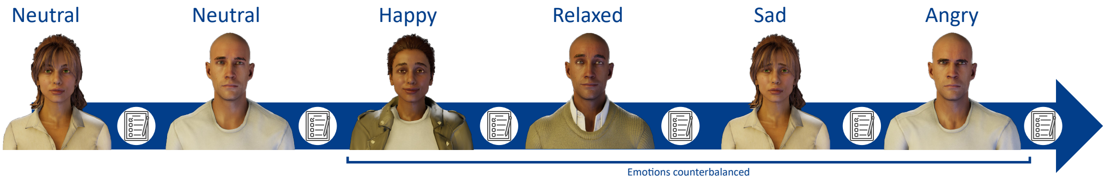


Figure 4.3: Scheme of the experimentation protocol

Following each conversation, participants were instructed to assess both the naturalness and realism of the social interaction. Naturalness is appraised in terms of the conversation's flow, spanning from "I felt very forced during the conversation" to "I felt very natural during the conversation". Realism assesses the content generated during the conversation, ranging from "It was an artificial conversation. It doesn't resemble a real conversation at all" to "It was a realistic conversation. The content was very similar to that of a real conversation". Both metrics employ a 7-point Likert scale.

In addition, they were asked to identify both the VH's emotions and their own, utilizing the Self-assessment Manikin (SAM) questionnaire [237]. This is a standardized measure of valence and arousal referring to Russell's Circumplex Model. In this context, participants were asked to rate both subjective and VH's valence and arousal using two 9-point Likert scales during their conversations. We hypothesize that a happy VH will induce high arousal and positive valence; relaxed VH, low arousal and positive valence; sad VH, low arousal and negative valence; and angry VH, high arousal and negative valence. We are setting 5 as the threshold on both scales, as it represents the midpoint of the Likert scale. Data collection was executed from October 2022 to February 2023.

4.2.8 Data analysis

The present analysis of the experiment is structured into five distinct sections. The first section focuses on evaluating the time-processing efficiency of the pipeline employed, paying particular attention to the time expended by the modules that incorporates AI-based APIs: audio transcription, LLM, and voice synthesizer. The pipeline log, based on UNIX time, was used for this computation, assessing the interval between input receipt and output provision by each module. The remaining pipeline modules, those not reliant on external AI-based API services, were collectively evaluated.

The second section embarks on a descriptive analysis of human-computer speech interaction, drawing upon features extracted from the conversations. For clarity, we define a 'conversation' as the total time frame from greeting to farewell. An 'interaction' refers to a cluster of one or more consecutive phrases that are conveyed concurrently within the same audio recording or synthesized by either the participant or the VH respectively. A 'sentence', on the other hand, pertains to a distinct phrase with an identified or fabricated endpoint. Our analysis covered the total duration of the conversation, the number of interactions per conversation, the number of sentences per conversation, and the number of words per conversation. Additionally, we evaluated the total duration and speaking duration per interaction. For human participants, interactions were dissected into three phases: (1) 'Time to start talking', representing the silent interval between the initiation of the recording and the beginning of speech, (2) 'Time talking per interaction', and (3) 'Time of final silence', the span between when participants stopped speaking and the conclusion of the recording. Silent periods were identified using pyannote voice activity detection model [236]. It should be noted that as the VH's voice is synthesized, there are no segments pertaining to 'Time to start talking' or 'Time of final silence'; hence, the total and speaking durations are identical in this instance. Lastly, the interaction time was further evaluated in relation to its position within the conversation for both the human interlocutor and the VH. This computation subsequently differentiated between the initial audio of the human and VH, the second audio, and so forth. Figure 4.7 clearly illustrates the average evolution of speaking duration throughout the conversation. Due to non-normality of the data based on Shapiro-Wilk test, Mann-Whitney-Wilcoxon test were applied to analyse differences between VH and subjects interactions.

As the final component of our analysis of human-computer speech interactions, we conducted a manual review to assess instances where errors in the interaction could be produced. These errors are separated in two, depending on the source, which are the silence detection module and the LLM exhibited errors. Errors arising from the silence detection pipeline were quantified by tallying the number of times in which the VH interrupted the subject during the subject's speech within each conversation. Conversely, errors stemming from incoherent responses generated by the LLM encompassed various categories, including repeated messages, statements made in the third person, messages that were inconsistent with prior conversation topics, the use of incorrect words, including foreign or nonexistent terms, and technical errors, such as instances where the VH failed to comprehend the human's sentence.

Naturalness and the realism of the conversations was analysed for each type of VH. A Friedman test is used to identify statistical differences among various VH conditions, averaging both neutral scores. This is followed by a post-hoc analysis using the Nemenyi test with Bonferroni correction to examine statistical differences between distinct states. Additionally, a Wilcoxon signed-rank test is employed to evaluate the average scores in neutral versus emotional conditions for each subject.

Finally, the emotional responses elicited in participants (i.e., the emotions experienced by the subjects), and the identification of the emotional state of the VH (i.e., the emotion the participant perceives the VH to be experiencing) have been explored in Section 3.4 and 3.5. The results were collected based on valence and arousal utilizing the SAM scale [238], employing a 9-point Likert scale. Once the normality of the distributions is verified, a repeated measures ANOVA is performed over the different emotional states. A post-hoc analysis with Bonferroni correction is performed to analyze statistical differences between the different emotional states. One subject was excluded from the analyses of naturalness, realism, emotion elicitation, and emotion identification because his/her score on the STAI questionnaire surpassed the 95$^{\text{th}}$ percentile (47 for women and 40-41 for men) [239].

4.3 Results

4.3.1 Technical performance

Figure 4.4 presents the distribution of processing times utilized by each module to generate an output. The LLM, specifically OpenAI's GPT-3, was the most time-consuming module with a mean processing time of 1.82 ± 1.46 s, constituting 41.07% of the total conversation pipeline duration. The 99^{th} percentile of the distribution is at 6.89 s. The audio transcription module, employing Google's Speech-to-Text technology, also demonstrated a substantial processing duration of 1.70 ± 0.62 s, corresponding to 38.14% of the total pipeline time. In comparison, the voice synthesizer module, powered by AWS Polly, required a relatively minor average processing time of 0.09 ± 0.04 s to complete the voice synthesis, amounting to only 2.13% of the pipeline's overall time. The remaining modules, encompassing data recording and storage as well as text cleaning, required an average processing time of 0.82 ± 0.42 s, equating to 18.65% of the total pipeline time. Overall, the entire conversation pipeline commanded an average processing time of 4.44 ± 1.77 s, where the 99^{th} percentile of the time consumption is located in 10.18 s.

4.3.2 Human-machine interaction

The average duration of a conversation is 3.44 ± 0.97 minutes, involving an average of 20.81 ± 6.51 interactions per conversation between the VH and the participant. Figure 4.5 shows the time distribution of the conversation durations. It illustrates two distinct distributions, separated by a temporal threshold set at 4 minutes to conclude the conversations. Conversations concluding below this threshold exhibit an average duration of 2.730 ± 0.712 minutes and includes 57.34% of the conversations. Conversely, the second distribution encompasses conversations that exceed the 4-minute threshold, terminating with an average time of 4.349 ± 0.163 minutes. Notably, 42.66% of conversations extend beyond this threshold.

The mean number of sentences used by the subject was 17.23 ± 5.60 sentences per conversation. In contrast, the VH uses significantly more sentences, averaging 23.71 ± 6.97. According to the Mann-Whitney test, there is a statistically significant difference (p-value < 0.001) between these two distributions. As for the word count, a conversation involves an average of 112.46 ± 54.85 words contributed by the human participant, while the VH uses more words, averaging 145.25 ± 51.46. Again, the Mann-Whitney test reveals a statistically significant difference (p-value < 0.001) between these distributions.

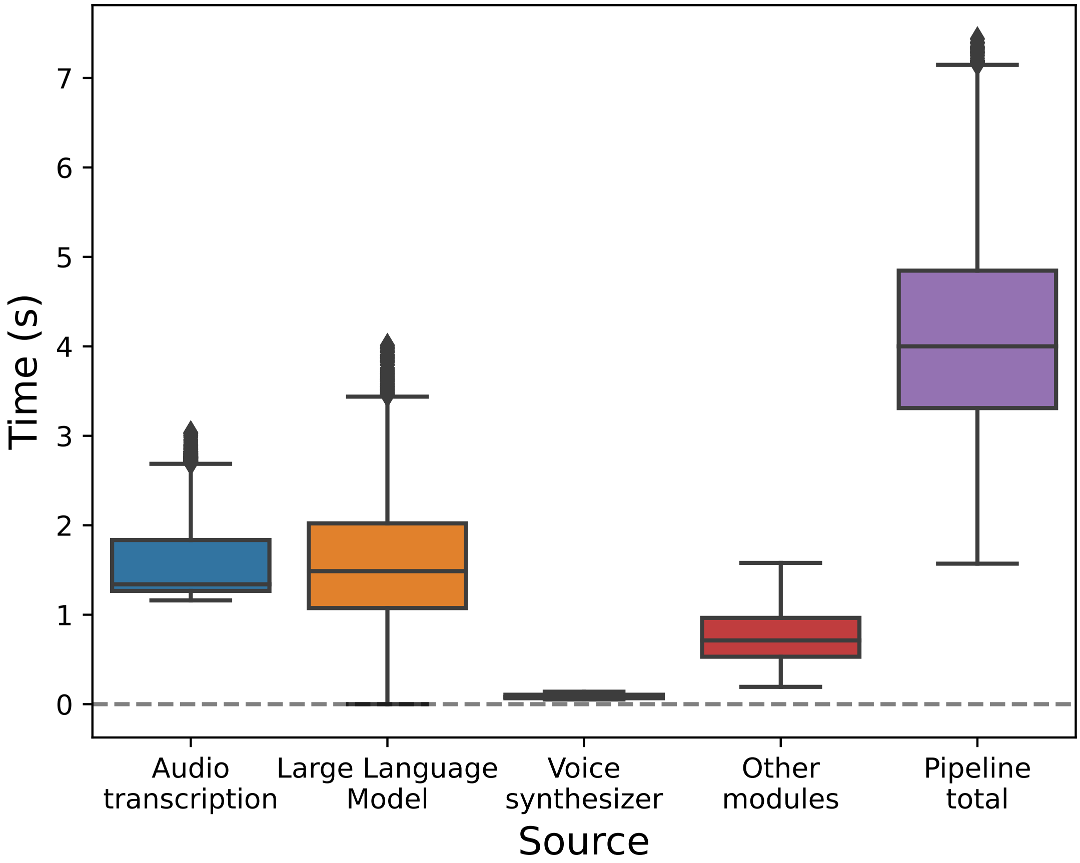


Figure 4.4: Boxplot of the time spent by the whole pipeline and splitted by each module.

With respect to audio interaction features, the extracted results are outlined below and visualized schematically in Figure 4.6. The human participant typically takes an average of 1.27 ± 1.91 seconds to begin speaking, with their speaking duration averaging 3.94 ± 1.61 seconds. The VH, conversely, engages in lengthier speaking durations, averaging 5.74 ± 3.85 seconds. A statistically significant difference is indicated by the Mann-Whitney test (p-value < 0.001). The silence detection pipeline's average time to cease recording following the participant's discontinuation of speech is 2.80 ± 0.54 seconds. The average audio duration for the human participant is 8.01 ± 3.66 seconds, while the VH's average audio duration is slightly lower, at 5.74 ± 3.85 seconds, presenting statistical differences (p-value < 0.001). The audio duration and the end of the VH time speaking is the same because the API of the module returns the audio with only the sentence of the VH. The use of the silence detection algorithm is not necessary.

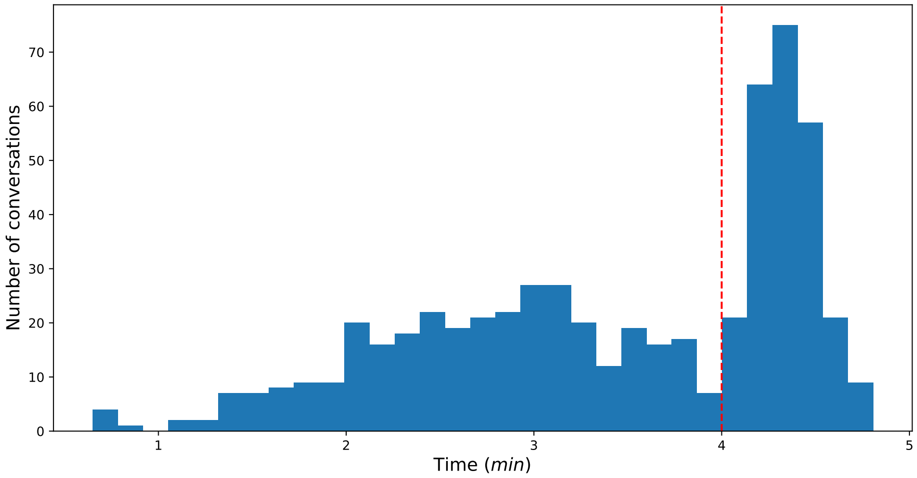


Figure 4.5: Distribution of the time conversation duration for all conversations.

Figure 4.7 illustrates the average speaking duration per audio segment, categorized by their position within the conversation for both the human participant and the VH. On average, the VH spends more time speaking than the human. However, similar trajectories are observed for both, as the initial and concluding audio segments tend to be shorter, while those in the third and fourth positions are typically longer.

Table 4.4 presents the percentages of errors per conversation and the corresponding impact on individual interactions arising from various error types, as well as the cumulative effect across all conversations and interactions. Among these error categories, the most prevalent error source stems from VH interruptions induced by the silence detection algorithm, affecting 1.22% of individual subject interactions, which form part of a total of 11.24% conversations. Concerning the LLM, inconsistent messages are the most common error affecting the 0.85% of the interaction, that form part of the 9.12% of the conversations. It is follow by technical errors, repeated message, 3^{rd} person message and incorrect words. A total of 2.62% of samples and the 23.28% of conversations have been affected by any different error from the LLM. In summary, the 3.84% of the interaction show some type of error, affecting 29.02% of the conversations.

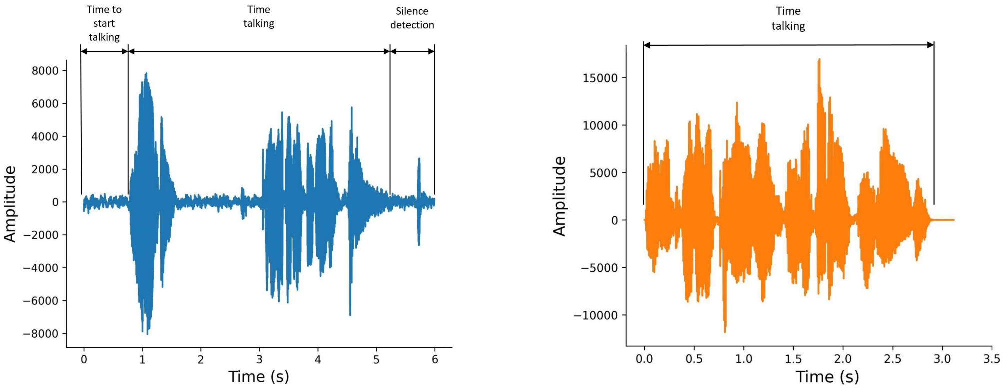


Figure 4.6: Schematic of the audio interaction features extracted for human and VH.

Table 4.4: Percentage of conversations and interactions affected by each type of error.

Source	Error type	Affected interactions (%)	Affected conversations (%)
Silence detector	VH interruption	1.22	11.24
LLM	Repeated message	0.59	6.35
	3^{th} Person message	0.33	3.59
	Inconsistent message	0.85	9.12
	Incorrect words	0.08	0.83
	Technical errors	0.77	8.29
	Total	2.62	23.28
Total affected		3.84	29.02

4.3.3 Naturalness and realism

The perceived naturalness (pertaining to conversation flow) and realism (related to conversation content) are analyzed based on the emotional states of the VHs. The averaged scores for each VH, along with the statistical values from the Friedman test across different groups, are presented in Table 4.5. This analysis reveals distinct differences in both naturalness and realism across various VHs. Post-hoc analysis, however, does not reveal any statistically significant differences between the four non-neutral emotions (relaxed, happy, sad, and angry). Yet, these emotions collectively exhibit statistical differences when compared to the neutral state in all cases, with the exception of the relaxed-neutral comparison in terms of realism. To further analyse these findings, we

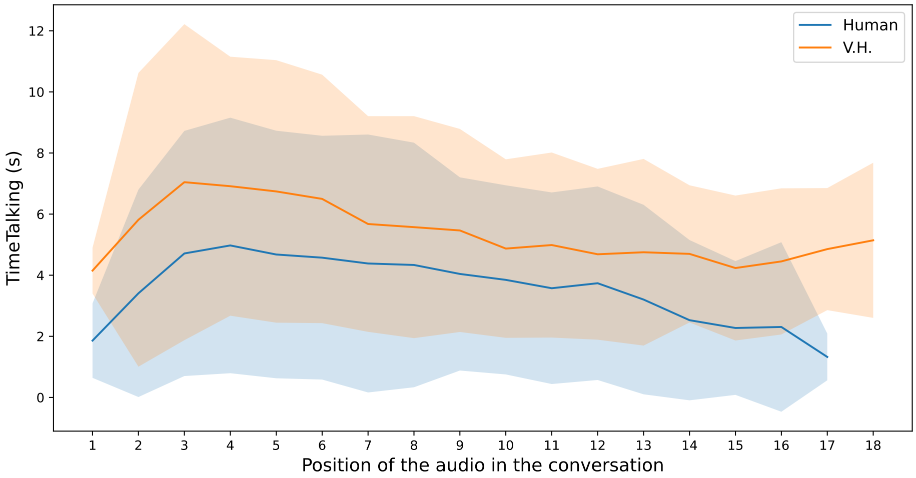


Figure 4.7: Mean duration of the audios in terms of each position in the conversation for the Human and the VH. The average time is represented by a line whereas the standard deviation of the duration is the shadowed area.

examine the statistical differences between the emotional and neutral states of the VHs. The Wilcoxon signed rank test reveal a statistically significant difference between both groups, with emotional states outperforming neutral ones by approximately one point in both naturalness and realism. The overall scores for naturalness and realism are 4.52 and 4.48 respectively, placing them above the midpoint on the Likert scale from 1 to 7, suggesting a general sense of satisfactory realism and flow in the conversations with the VHs.

4.3.4 Emotion elicitation

Table 4.6 presents the averaged score of participant self-assessment using SAM for valence and arousal using the Likert scale from 1 to 9. These scores describes the VH's ability to elicit emotions in subjects. Repeated measures ANOVA reveals statistical differences between conditions in terms of valence, but not in terms of arousal. As illustrated in Figure 4.8, the positions of both the angry and relaxed VHs align with their respective theoretical emotion quadrants according to Russell ([4]). However, the VH exhibiting happiness induces less arousal than expected, resulting in its placement outside of its corresponding quadrant. Conversely, the VH expressing sadness elicits a greater degree of valence than hypothesized, leading to its repositioning into the relaxed

Table 4.5: Average subject assessment of the VHs naturalness and realism, standard deviation in parenthesis

Personality	Naturalness (1-7)	Realism (1-7)
Neutral	4.16 (1.32)	4.20 (1.38)
Angry	4.64 (1.70)	4.75 (1.70)
Happy	5.00 (1.58)	4.74 (1.65)
Sad	5.02 (1.72)	4.89 (1.60)
Relaxed	4.90 (1.75)	4.64 (1.45)
p-value	< 0.001 (***)	< 0.001 (***)
Neutral	4.16 (1.32)	4.20 (1.38)
Emotional	4.89 (1.39)	4.75 (1.30)
p-value	< 0.001 (***)	< 0.001 (***)
All	4.52 (1.41)	4.48 (1.37)

quadrant. A pairwise comparison was undertaken to examine the distribution of arousal and valence experienced by participants after interacting with the VH, depending on each distinct emotional state of the VH. The results of post-hoc analysis are displayed in Table 4.7. In regard to valence, the hypothesized differences manifested in pairwise assessments. These include contrasting anger and happiness, anger and relaxation, along with happiness and sadness. However, the comparison between sadness and relaxation was the sole exception, displaying no notable differences. On the other hand, no statistical differences were identified in terms of arousal elicited by the different emotional states of the VHs.

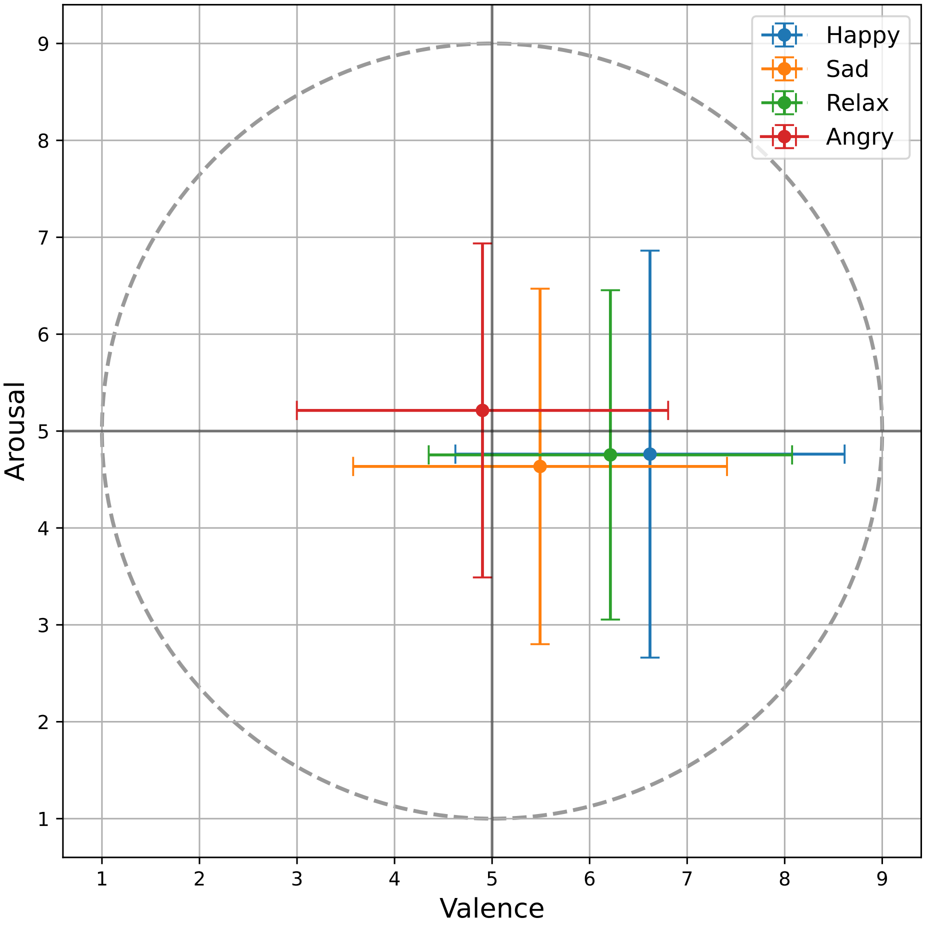


Figure 4.8: SAM questionnaire results for subject self-assessment. The dot shows the average value of the valence and arousal. The vertical and horizontal lines show the standard deviation in terms of arousal and valence for each different emotion.

4.3.5 Emotion identification

The average score for subject emotion identification of the VH is shown in Table 4.8. The results of the repeated measures ANOVA show statistical differences in valence and arousal. Figure 4.9 shows that the location of the relaxed VH fits the theoretical model

Table 4.6: Scores of the valence and arousal for subject self-assessment. Standard deviation is inside parenthesis. (*) p-value < 0.05, (**) p-value < 0.01, (***) p-value $<$ 0.001.

Personality	Valence (1-9)	Arousal (1-9)
Angry	4.89 (1.89)	5.23 (1.72)
Happy	6.62 (1.90)	4.66 (2.02)
Sad	5.39 (1.93)	4.69 (1.79)
Relaxed	6.16 (1.89)	4.74 (1.67)
p-value	< 0.001 (***)	0.262

Table 4.7: Statistical comparison between the different emotional states for subject self-assessment over the different VH personalities. (*) p-value < 0.05, (**) p-value < 0.01, (***) p-value < 0.001.

Personality		p-value	
		Valence	Arousal
Angry	Happy	< 0.001 (***)	0.078
Angry	Sad	0.118	0.053
Angry	Relaxed	< 0.001 (***)	0.060
Happy	Sad	< 0.001 (***)	0.890
Happy	Relaxed	0.148	0.715
Sad	Relaxed	0.014 (*)	0.827

of arousal and valence [4]. Angry VH is very close to its theoretical quadrant, with a high arousal and a valence close to the scale neutral point. Similarly, happy VH is very close to its corresponding with a positive valence and an arousal close to the scale neutral point. Finally, the sad VH gets slightly more arousal and valence than expected, achieving the position in the quadrant correspondent to the happy VH. Table 4.9 includes the post-hoc analysis. Regarding valence, the study successfully demonstrated all the hypothesized differences, specifically between the emotional states of anger and happiness, anger and relaxation, happiness and sadness, as well as sadness and relaxation. As for arousal, not only did the study confirm hypothesized contrasts, such as those between anger and sadness, and anger and relaxation, it also unveiled unanticipated differences, namely between anger and happiness, and sadness and relaxation. However, the results did not corroborate hypothesized distinctions in the case of happiness versus sadness and happiness versus relaxation.

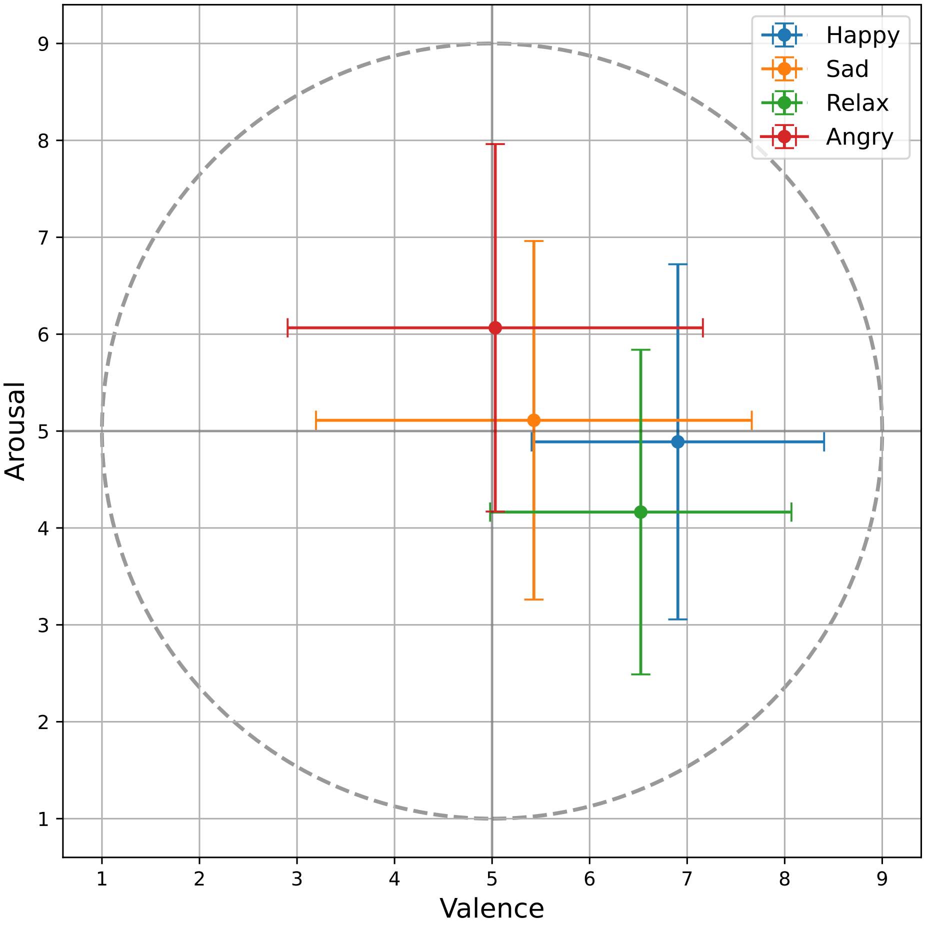


Figure 4.9: SAM questionnaire results for subject identification. The dot shows the average value of valence and arousal. The vertical and horizontal lines show the standard deviation in terms of arousal and valence for each different emotion.

4.4 Discussion

This work introduces the first comprehensive VH system specifically designed for real-time emotion elicitation during semi-guided conversations. It was created through the integration of cutting-edge AI and VR platforms, based on a cognitively-inspired archi-

Table 4.8: Scores of the valence and arousal for subject identification. Standard deviation is inside parenthesis. (*) p-value < 0.05, (**) p-value < 0.01, (***) p-value < 0.001.

Personality	Valence (1-9)	Arousal (1-9)
Angry	4.98 (2.15)	6.15 (1.90)
Happy	6.87 (1.54)	4.91 (1.83)
Sad	5.39 (2.29)	5.08 (1.85)
Relaxed	6.51 (1.54)	4.16 (1.66)
p-value	< 0.001 (***)	< 0.001 (***)

Table 4.9: Statistical comparison between the different emotional states for subject self-assessment over the different VH personalities. (*) p-value < 0.05, (**) p-value < 0.01, (***) p-value < 0.001.

Personality		p-value	
		Valence	Arousal
Angry	Happy	< 0.001 (***)	< 0.001 (***)
Angry	Sad	0.285	0.001 (**)
Angry	Relaxed	< 0.001 (***)	< 0.001 (***)
Happy	Sad	< 0.001 (***)	0.493
Happy	Relaxed	0.119	0.002 (**)
Sad	Relaxed	< 0.001 (***)	0.004 (**)

tecture where each module aims to perform a particular task in a mind/body scheme. The primary objective of this research was to validate the developed VHs as emotional stimuli, exploring the range of emotions evoked in the subject during their interaction with the VH. The generation of the emotions was based on modifications in VH's body (i.e. facial expressions) and the mind (i.e., attitudes, moods, and motivations represented in the prompts sent to the LLM). For this purpose, a total of twenty VHs were crafted, having five emotional states (neutral, anger, happiness, sadness, and relaxation), two genders (male and female), and two types of appearance (casual and semi-formal).

4.4.1 Virtual Human architecture

This research introduces a novel framework for developing VH, founded on the idea of bifurcating modules into mind and body, a concept adapted from the architecture proposed by [220]. The body system is further delineated into action and sensory subsystems. The action subsystem is engineered to generate real-time audio-visual content for the VH, thereby facilitating real-time interactive conversations. This involves the integration of a lifelike avatar completed with facial expressions, synthesized voice, and lip synchronization. Notably, we leveraged AI-based models: AWS Polly API for instantaneous voice synthesis, and Nvidia audio2face for real-time lip synchronization. This approach culminates in a realistic audio-visual interface that is projected on a large screen, displaying a life-size avatar. The sensory subsystem equips the VH with auditory capabilities. It comprises a silence detection module, which conducts ongoing analysis to ascertain when the interlocutor has finished speaking, and an audio transcription module based on Google's AI model. This module provides the cognitive competency to interpret the spoken audio. The mind system furnishes the VH with psychological traits along with cognitive and emotional aptitudes. Specifically, we endowed it with cognitive proficiencies such as language comprehension, rational thought, creativity, and memory, utilizing GPT-3. Additionally, we incorporated psychological traits and states such as life-history, context, attitudes, moods, and motivations into the prompts sent to the LLM. To ensure temporal coherence, we integrated a memory module. The participants were equipped with a microphone enabling them to converse with the VH. Further, data pertaining to ET, EEG, ECG, and EDA were collected. These biometric data will fuel future research, facilitating the development of automatic emotion recognition systems and biofeedback intelligence. Several recent studies have deliberated on approaches to create VHs. For instance, [240] proposed a four-class taxonomy rooted in the degree of form and behavioral realism. A 'Digital Human' necessitates a realistic anthropomorphic appearance, intelligence, autonomous and natural verbal and non-verbal communication, cognitive, affective, and social abilities. Our present architecture attempts to encompass all these elements, thereby illuminating the path for future research in the evolution of VH.

4.4.2 Technical performance

The architecture under study has been technically dissected, with time expenditure for each module thoroughly analyzed. Initial focus was directed towards three modules that incorporated AI-based API services, namely the voice synthesizer, the audio transcription, and the LLM. Remaining modules were evaluated collectively. The LLM module emerged as the most time-consuming, with an average time expenditure of $1.826 \pm 1.462\ s$ (41.07%). While alternative LLMs offering greater efficiency are available, GPT-3 was chosen for its superior conversation quality and coherence compared to other text generator models. Moreover, the usage of GPT-3 through OpenAI's API bypasses the need for local computation of model responses, thus preventing potential tool slowdowns. At the onset of data collection, GPT-4 was not available. Importantly, the selected model should balance between high-quality text generation and time efficiency to maintain conversational naturalness. Slow models may compromise this aspect. The second most time-consuming module is the audio transcription module, consuming an average time of 1.70 ± 0.62 s (38.14%). The Whisper ASR model was considered but was ruled out due to its local computation requirement and time-intensive nature. The voice synthesizer module operates the fastest AI-based API for generating the VH's voice, requiring 0.09 ± 0.04 s (2.13%). Although synthetic speech generation models are emerging in the open-source realm, as far our knowledge, the AWS's speed currently outperforms those deployed locally. As new AI models continue to evolve and garner interest, it is anticipated that more accurate and efficient models will be developed and become accessible. Currently, reliance on external AI-based APIs is essential to yield high-quality VHs, primarily due to (1) the scarcity of open-source AI models for translate and synthesize voices, as well as generate text, and (2) the high computational resources demanded by such AI models, necessitating large-scale computing distributions. Nevertheless, the modular nature of the proposed VH architecture facilitates the independent testing and time optimization of new AI models, enhancing each module individually. The remaining modules collectively consume 0.82 s, accounting for 18.650% of the total communicative pipeline. Despite its relatively small proportion, this time duration could potentially be curtailed. Further exploration into alternative data processing or saving mechanisms could reduce this duration, thus contributing to a more fluid conversation.

4.4.3 Human-machine interaction

Descriptive features were extracted from audio recordings of each conversation, providing a quantitative evaluation of duration and the nature of verbal human-machine interactions between the subject and the VH. The conversation's average duration amounted to 3.38 ± 0.98 minutes, encompassing an average of 20.61 ± 6.40 interactions. However, the histogram (Fig. 4.5) clearly displays two normal distributions with distinct central tendencies. 57.34% of the conversations belong to the distribution that does not exceed the 4-minutes threshold, with a mean of 2.730 minutes. Notably, 42.66% of the conversations exceed the threshold. This indicates high engagement experienced by the subject during the conversation and suggests a considerable amount of information condensed within the conversations.

On the other hand, statistical analysis unveiled that the VH consistently employed on average more sentences (23.71 vs 17.23, +37%) and words (144 vs 110, +30%) than human participants, culminating in extended periods of VH's speech (5.74 vs 3.94 s, +45%). These results indicate a higher level of dominance by the VH in the conversation, likely attributable to its design focus on triggering emotional narratives. Given this, the engagement appears quite balanced, with users demonstrating a high level of involvement during the conversations, as evidenced by sentence and word counts. Additionally, the duration of interactions suggests that the conversations were notably fluid. Maintaining such fluidity and balance was crucial in this experimental protocol to prevent excessive dominance by either the VH or the subject in conversations. It's pertinent to note that while subjects' speech rates may vary, the VH maintains a fixed speech rate which could differ from that of the subjects. Future research will examine variances in semantic and prosodic speech features. Furthermore, the pursuit of developing more sophisticated VHs that can closely mimic human speech remains an ongoing challenge.

The audio derived from the interaction with the subject can be segmented into three parts: (1) time to start speaking, (2) speaking duration, and (3) silence detection. Subjects, on average, commenced speaking 1.27 s after recording initiation. This part represents a natural waiting time until the subject detects that the VH stops speaking. However, the silence detection module holds paramount importance within the pipeline, tasked with the precise identification of dialogue termination or silence within audio streaming. However, a potent AI model capable of detecting the end of a dialogue in streaming audio is yet to be realized. To bridge this gap, we implemented a silence detection algorithm based on audio activity detection, facilitating uninterrupted dialogues without preset time thresholds for speech pauses. Nevertheless, the algorithm necessitates audio recording each time it's invoked, potentially causing considerable conversation delays. The average time spent during an interaction loop, identifying when a user concludes speaking, is measured to be $2.808 \pm 0.546\ s$. This duration is amenable to optimization for delay reduction. Our algorithm employs diverse time thresholds to safeguard against information loss. Humans can quickly tell when someone has finished speaking due to the rhythm and tone of speech, or what's known as prosody. This suggests that upcoming silence detection systems should use these prosodic features to more quickly detect when speech has ended instead of a fixed threshold of activity detection. It should be noted that the pipeline duration, from the moment the silence detection module concludes that the subject has ceased speaking until the VH begins its response, averages at 4.445 s (Fig. 4.4). Comparatively, the real-time silence detection requires 2.808 s. Thus, a substantial portion of the waiting period for the subject, prior to receiving a response, is dedicated to ascertaining that the subject has completed their dialogue.

The errors arising from human-machine interaction were manually reviewed post-experimentation and categorized into distinct types, such as VH interruptions, repetitive VH messages, and incorrect word usage, with the corresponding percentages of affected interactions and conversations detailed in Table 4.4. Although the silence detector interrupts only 1.22% of the interactions, it can disrupt the flow of 11.24% of them. Consequently, it is a critical module slated for future enhancements, incorporating a more sophisticated model considering not just voice activity but also prosodic dynamics to discern if the subject has concluded the interaction. In contrast, the LLM issues affect between 0.08% and 0.85% of interactions, impacting a total of 23.28% of conversations. Utilizing future versions of LLM will ameliorate these ratios. Nonetheless, it is crucial to highlight that while these issues impede a specific interaction during the conversation, they don't necessarily affect the entire conversation. The subjects' perception of the conversation flow is analyzed in the subsequent subsection.

4.4.4 Naturalness and realism

The naturalness and realism of the model has an average score of 4.52 and 4.48 over 7 correspondingly. The highest score in naturalness and realism is achieved by the sad VH which is 5.02 and 4.89 respectively. On the other hand, it can be seen how the neutral VH achieves, in average, lower punctuation's in both scores. This could be due to different reasons. The first one is that emotion increases the sense of presence, as stated in previous literature [78,241]. Therefore, as long as the emotion is expressed by the VH, the engagement and social presence of the subject increases, enhancing the naturalness and realism of the conversations. This results present a new step in analysing the relationship between presence and emotions during conversations with VH, supporting previous evidence. However, we need to consider that the neutral VHs are the ones that are seeing firstly by the subject, since they were used as an adaptation process, allowing subjects to familiarize with the system. As the experimentation progress, they get used to the look and the interaction with the VH due to their exposure, and the scores of realism and naturalness may increase.

4.4.5 Emotion elicitation

The results of the SAM regarding the self-asesessment showed differences in terms of the valence elicited in the subjects by VHs (Table 4.6). In general, there is a VH's tendency of eliciting emotions with positive valence during the conversations, since the lowest score (4.89), achieved by the angry VH, is very closed to the the scale neutral point of 5. Nevertheless, we achieved all the expected differences, acomplishing the goal of eliciting varying emotions in terms of valence. To reduce the positive bias, stronger negative narratives may be designed, and the VH may be improved to increase the subject's empathy. In terms of arousal, the subjects did not report any differences. Therefore, we were not able to elicit coherent arousal reactions in subjects. One potential reason is that VHs did not change the prosody features of the voice depending on the emotional state such as volume, tone and speech rate, and it is well-known that para-verbal aspects of human communication carries on a significant portion of the emotional content expressed by individuals [242]. This could affect the arousal perception as well as its elicitation. Further research is needed to elicit arousal during conversation with VHs and to explore its relationship with para-verbal features.

4.4.6 Emotion identification

In the case of emotional identification of the VHs, the expected differences were all achieved in valence. Therefore, the subjects were able to identify all the emotions designed in terms of valence. In the case of arousal, several differences are shown, supporting the hypothesis that although coherent arousal was not elicited in the subjects, they were able to identify different arousal patterns between VHs. In general, the negative emotions have been identified as higher arousal than the positive ones. It could be related to the negative bias in social-emotional interaction, since humans use negative information far more than positive information [243]. On the other hand, happy VH presented lower perceived arousal than expected as it was close to the scale neutral point, and further efforts are needed either to elicit effusive happy states, or to reduce the negativity bias effect.

4.4.7 Future of virtual humans

Recent reviews underscore the crucial role of ecological and social emotion elicitation methods in various fields, particularly healthcare, including the assessment, monitoring, and intervention of mental health disorders like depression, anxiety, bipolar, and psychotic disorders. [244] highlighted the urgent need for comprehensive evaluation of diverse healthcare conversational agents, focusing on their acceptability, safety, and effectiveness. Their review revealed that most conversational agents described in the literature are text-based, machine learning-driven, and delivered via mobile apps. [245] emphasized the significance of recent technological advancements in chatbots to mental health research and care. They detailed how smartphones, social media, artificial intelligence, and VR offer new opportunities for "digital phenotyping" and remote intervention, but noted the need for further efforts towards solidifying implementation. [246] proposed an empathic conversational agent framework, combining natural language processing techniques and artificial intelligence algorithms for real-time monitoring and co-facilitation of patient-centered healthcare. However, their system relies on a rule-based emotional message generation, limiting the potential for open-ended conversation. Furthermore, they did not integrate the conversational agent with in an avatar. In a comprehensive review of emotional chatbots, [247] combines VR, a chatbot and a voice user interface in a smartphone aplication, but without a Large Language Model. [248] discussed the current limitations of the state-of-the-art. They pointed out that generative chatbots are still not based on LLM, possibly due to their novelty. Moreover, no research thus far has incorporated a conversational agent into an avatar.

Denecke et al. [249] presented the most recent review of Conversational Agents in health, illustrating that the current landscape of health CAs is predominantly characterized by rule-based, simple systems in terms of CA personality and interaction.

In other fields, [250] developed a VR scenario with a chatbot to facilitate job interview training. Face recognition and sentiment analysis were utilized to analyse and provide feedback to the participant. [251] provided a review of chatbots in education, outlining potential future areas of education that could benefit from this advanced AI technology. However, they have not yet presented any comprehensive VH.

Despite the rise of AI-based technologies in fields such as natural language processing and computer graphics, there remains a lack of studies in affective computing that analyze human emotional responses using these methods. There is no existing work that integrates advanced technologies like audio transcription, LLM, speech synthesis, lip synchronization, and VR to create a realistic VH capable of conducting real-time, open-ended conversations. The system developed in this study represents a significant advancement by merging these technologies.

Recent reviews of chatbots, conversational agents, or VHs suggest they will have significant implications in several fields such as health and education. It is our hope that the present work can guide future applications of VH in these areas.

4.4.8 Limitations and future research

The present study has certain limitations that future research should aim to address. A significant constraint lies in the VR platform employed, which hampers the natural movements and gestures of the VH, diminishing its realism. The platform does not accommodate hand or leg movements, and the VH's blinking pattern was preprogrammed, limiting the organic behavior. Furthermore, the absence of an effective AI model to control voice modulation, rhythm, and tone could influence the participant's empathetic response towards the VH. The static facial expression of the VH during the conversation also reduces the naturalism and realism of the interaction. Incorporating dynamic changes in the VH's facial state throughout the conversation could greatly enhance its lifelike quality. Another area of potential improvement pertains to the use of advanced language models, such as GPT-4, which could potentially provide superior text generation performance, thus improving the natural flow of conversation between the participant and the VH. Upgrades in audio transcription models could further enhance the realism of the VH's conversation. Looking forward, future research should concentrate on refining the VH to serve as a more ecologically valid social environment for eliciting psychological phenomena via emotions. Such efforts will aid in expanding the potential applications of VH across various disciplines. As we chart the course of our future research trajectory, we plan to expand our scope by developing emotion recognition systems grounded in biometric data already collected, including ET, EEG, ECG and EDA. Additionally, we aim to introduce a new phase in our experimental design by including subjects exhibiting depressive symptoms. Our objective is to architect an automated health assessment and monitoring system driven by supervised machine learning. This system will leverage our ecologically valid environment during emotional elicitation. Our hypothesis proposes that such systems will instigate the manifestation of specific phenomena tied to depressive symptoms. This is based on the understanding that depressive symptoms often correlate with impaired social skills and amplified negative emotional states, as supported by prior studies [213,252]. Therefore, our approach aims to leverage these insights, advancing the development and application of intelligent systems for improved mental health diagnosis and treatment. In addition, further research needs to assess the influence of VH gender on emotion elicitation, as well as whether there are subject gender differences in terms of assessing the system.

4.5 Conclusion

The findings presented here mark a new step in the field of AfC and its applications, presenting the first real-time conversational VH capable of engaging in semi-guided dialogue based on LLM. Drawing on a modular cognitive framework, we divided the VH into distinct mind and body components, providing a realistic life-size avatar with language comprehension, rational thought, creativity, and memory. We presented an extensive validation of the system in terms of technical performance, human-machine interaction, naturalness and realism, and emotion elicitation and identification. This breakthrough marks a considerable progression in the crafting of ecologically valid social virtual environments and the ability to generate emotion-driven responses that are shaped by the nuances of social and contextual factors. The intelligent system unveils a multitude of possibilities and potential applications, spanning mental health and permeating a plethora of other disciplines.

Acknowledgements

Thanks to Pepe Roda and Eduardo Manzano for modelling the avatars, and Luciano Spinella, Carmen Calero and Marina Catala for assisting the experiment.

Model availability

The conversational AI pipeline can be found in the following link
https://github.com/ASAPLableni/LableniBOT.

Chapter 5

Emotion and depression recognition through conversational Virtual Humans

Abstract

The investigation of emotion recognition (ER) through virtual reality (VR) has garnered significant attention in recent years. This tool could be extended to routinely environments to examine the emotion elicitation. This study proposes an experimentation that involves four virtual human (VH), each designed to convey distinct emotional states, to elicit different emotions in individuals. Furthermore, it is studied the recognition of a mental illness such as depression. To achieve this, a study was performed engaging a total of 98 participants in conversations with the VHs, including eye-tracking (ET) and electrodermal activity (EDA) signal analysis. The ET and EDA signals were independently processed. From the ET data has been obtained fixations, saccades and blinks, while specific areas of interest (AoI) were defined for the VHs, generating distinctive features. On the other hand, the EDA signals underwent thorough pre-processing, including artifact correction and signal decomposition into phasic and tonic components. Machine learning (ML) algorithms are finally used to perform classification tasks over different targets. Notably, our research yielded promising results in the recognition of depression during VH interactions, achieving a balanced accuracy rate of 0.685. However, the performance of ER varied across different targets, with the more robust result obtained for the classification of VH valence, reaching a balanced accuracy of 0.655. This

research marks an initial exploration of AI-based VH applications in the domains of ER and depression assessment in VR, offering a glimpse of their potential utility in diverse fields such as healthcare, education, and even the realm of interactive entertainment, including video games.

5.1 Introduction

Affective computing (AfC) encompasses various approaches, with one of the most significant being Russell's circumplex model [4]. This model characterizes emotions in a 2D affective space defined by two axes: arousal and valence. Russell's model provides a numerical description for different emotions within this 2D framework. Emotion recognition (ER) is the field dedicated to the identification of human emotions, and it represents a relatively new and evolving area of research where technology plays a pivotal role. In the course of studying this field, various techniques have been developed, drawing from disciplines such as computer science and statistical analysis. Much of the effort in this field is focused on automating ER through methods like signal processing, facial recognition, or speech recognition, all geared toward precise ER. Additionally, the accuracy of ER significantly improves when various techniques are combined, utilizing data encompassing text, biosignals, audio, and video. Numerous works have delved deeply into ER using diverse techniques. For instance, the study by Lim et al. [253] introduces a model that incorporates convolutional and recurrent neural networks for speech ER, achieving a high level of accuracy in the recognition task. Ghofrani et al.'s work [254] demonstrates that real-time ER through facial analysis is now feasible with a high degree of accuracy. Furthermore, Cong et al.'s research [255] illustrates how various signal processing techniques, such as heart rate variability (HRV) and electrodermal activity (EDA), can enhance ER, even when employing signals that have already undergone extensive study and development.

Eye-tracking (ET) stands as one of the fundamental physiological indicators for measuring visual attention. This technology not only tracks the subject's gaze but also captures additional parameters like blink frequency and pupil diameter [148]. Previous research, such as the study by Tarnowski et al. [256], has explored ER tasks by extracting features from eye movements, including fixations, saccades, and pupil diameter. Notably, they achieved a remarkable maximum accuracy of 80% using a support vector classifier (SVM). Furthermore, the examination of visual attention through the concept of areas of interest (AoI) has proven to be a valuable approach in assessing task performance.

For instance, the research conducted by Houwei et al. [257] defines various AoIs, where fixation time and other features were meticulously analyzed. This feature extraction methodology led to an accuracy of 84.1% in recognizing three distinct emotional states: positive, negative, and neutral. Moreover, novel techniques, such as deep learning (DL), have been applied to analyze ET data, as demonstrated in the work by Aracena et al. [258]. In this study, both the temporal information of gaze and pupil diameter were utilized to recognize emotions, yielding favorable results.

Another highly utilized signal in ER is EDA. EDA reflects the activity of the sympathetic nervous system, making it a crucial signal in ER studies. EDA can be decomposed into tonic and phasic components through de-convolution techniques [160, 259]. The tonic component is associated with slower movements and changes in the skin conductance, while the phasic component represents rapid movements of the signal, often referred to as the skin conductance response (SCR). The SCR commonly provides the features used in EDA-based studies, offering valuable information for a wide range of scientific research fields [161]. This approach has received significant attention from psychology and health-related studies [162]. In clinical analysis, SCR is used to assess pain, stress, schizophrenia, and peripheral neuropathy [159, 160]. According to Sharma et al. [260], the galvanic skin response (GSR) obtained through the EDA signal can assess stress and anxiety in patients and can also be used for behavioral therapy. This signal has been studied as a standalone biomarker in various research works. For example, the study by Ayata et al. [261] focused on ER, categorizing the arousal and valence of subjects using the EDA signal. In this work, the Deap dataset was employed, where subjects underwent various visual stimuli while their EDA signal was recorded. This model achieved an accuracy of 71.53% and 71.04% for arousal and valence, respectively, using a random forest classifier (RFC). Innovative approaches have also transformed the EDA signal into a spectrogram. The work by [262] used a 2D convolutional DL model to classify the EDA signal into four different emotions (amusing, boring, relaxing, and scaring), achieving a maximum accuracy of 80.20%. Furthermore, advancements such as feature extraction from frequency and time-frequency decomposition [263] have demonstrated high accuracy in arousal and valence recognition, reaching 85.75% and 83.90%, respectively. These research findings collectively demonstrate the diverse and complementary processing methodologies for the EDA signal. However, since changes in GSR signal often manifest slowly and exhibit a time delay, it is common practice to complement the study of EDA signal with other types of biosignals, such as HRV or even ET. This combination allows for a more comprehensive and accurate understanding of emotional responses and cognitive processes

The exploration of AfC requires the incorporation of highly realistic scenarios to elicit a profound sense of immersion in subjects, aiming to evoke emotions authentically. VR plays a crucial role in enabling the development of immersive environmental simulations, allowing users to engage as if they were in the real world [77]. These simulations span a wide variety of setups, encompassing different formats and platforms [264]. The progressive integration of VR technology into research has enhanced the sense of immersion in virtual environments (VE), a pivotal factor for achieving a lifelike experience. Immersion in VR is defined as the objective level of fidelity provided by a VR system and is inherently associated with the employed technology [265]. The advancements in VR technologies have significantly enriched research focused on understanding human behavior [128]. Beyond conventional self-assessment methods, VR offers the opportunity to integrate various implicit measures capturing unconscious processes, including EDA, HR [129, 136], and ET. Numerous studies have explored human behavior within VR environments, concurrently recording biosignals to discern patterns in interactions. For instance, de-Juan-Ripoll et al. [86] research investigates subjects' risk-taking behavior using ET and EDA signals, revealing distinctive patterns enabling the classification of subjects into high or low risk-taking categories. Similarly, Pinto et al. [266] explore various biosignals, including ECG and EDA, to assess emotions during video viewing, achieving high accuracy of 69.13% for arousal and 67.75% for valence in the ER task. Moreover, the work of Vicente-Querol et al. [267] compares ER in immersive and non-immersive VR settings, indicating a marginal improvement in the VR environment for emotion recognition. Collectively, these studies emphasize the growing significance of combining VR with biosignal recording to gain insights into human behavior and emotional responses.

On the other hand, there have also been significant advances in AI-based technology. For instance, individuals can interact and maintain a normal conversation through a language model facilitated by generative pre-trained transformers (GPT) models, such as GPT-3 by OpenAI [104, 268], representing a noteworthy advancement and emerging as one of the most sophisticated language models to date. Additionally, interactive chatbots like Alexa, Siri, or Cortana have already existed [223, 224]. Various models, such as speech transcription from Whisper [227] or Google's Speech to Text, are currently available and can achieve voice transcription with very good results not obtained previously. Moreover, platforms like AWS Polly have achieved commendable results in voice synthesis across various languages, offering a high level of flexibility. In summary, the current state of AI technology enables the enhancement of elements within virtual experiments. More complex yet realistic experiments can be developed using the technology mentioned above. The utilization of the latest AI-based technology in conjunction with VR allows for the introduction of elements that are unattainable in real-world experimentation, such as virtual humans (VH).

VHs are human-like characters that typically interact through computer screens and/or speakers. They exhibit human-like behaviors, including speech, gestures, as well as human characteristics like emotions, empathy, and memory [98]. The construction of a VH entails the utilization of various tools, such as automatic text generation, speech synthesis, and the interface through which the VH is presented. However, all these technologies could be used nowadays, allowing the creation of a VH. The use of these advances could replicate difficult dynamics, such as social interactions through a conversation, something that couldn't be achieved without this technology. The utilization of this technology fosters a more natural interaction, thereby achieving a more realistic elicitation of emotions.

Numerous studies have utilized VHs to explore subjects ability to recognize emotions. For instance, the work of Finkelstein et al. [269] introduces a virtual platform featuring VHs to facilitate the study and teaching of ER. This platform incorporates various interactive games to effectively demonstrate ER. In the study by Zibrek et al. [270], the focus is on assessing subjects capability to discern emotions conveyed by the VH through different gestures and facial expressions. Additionally, this study examines the correlation between VH gender detection and ER. Notably, the researchers found that the model used to display motion did not impact gender perception but did influence ER. In the work of Durupinar et al. [271], different VHs are employed for ER across various genders, ages, and races. Statistically significant effects emerged among different groups for individual emotion types. These studies collectively explore the subject's capacity to recognize emotions conveyed by VHs. However, it is worth noting that studies focusing on the elicitation of emotions in subjects during interactions with VHs are currently lacking. The work of Llanes et al. [272] studies emotion elicitation and recognition through the use of four VH with different emotional states and two neutral VHs. The results show statistical differences in most of the four emotions examined. However, there is a dearth of research that investigates ER in the interaction with VHs through the information collected by biosignals.

The amalgamation of various technologies described not only has the capability to recognize social emotions but can also identify more emotional disorders, including those classified as mental illnesses such as depression. Depression, which affects 264 million people globally, has a significant impact on the quality of life and can even lead to severe outcomes, including suicide [273]. Given that traditional self-reported assessments are susceptible to biases and may not accurately reflect real-world scenarios, it is imperative for primary care centers to augment clinical assessments with digital screening tools. These tools can facilitate early diagnosis, classification of symptom severity, and appropriate referrals for individuals struggling with depression. The advancements in digital technology present opportunities for monitoring cognitive and behavioral development, thereby contributing to precision medicine in the field of mental health diagnosis. This approach entails the identification of valid biomarkers and behavioral indicators, which, in turn, enable the development of tailored preventive and treatment interventions. These interventions can be customized to suit individuals' unique characteristics and needs throughout their lifespan. This holistic approach represents a promising avenue for improving mental health diagnosis and patient care.

There is compelling evidence that ET serves as a suitable biomarker for correlating with depression, primarily through the measurement of visual attention [274]. ET technology has demonstrated the capacity to detect gaze patterns associated with depressive symptoms by comparing eye gaze behavior towards various stimuli between control subjects and individuals exhibiting depressive symptoms [275]. Additionally, the EDA signal has been investigated for recognizing depression symptoms. Kim et al.'s work [276] demonstrated an accuracy of 74% and a sensitivity of 74% in recognizing depressive subjects using only EDA features from stress and relaxation tasks. However, the study of this signal is commonly conducted in conjunction with other biosignals, such as the electrocardiogram (ECG) signal. For example, the work of Ding et al. [277] delves into the recognition of depression patients using a combination of EEG, EDA, and ET. The study revealed that the integration of data from different sources enhances the performance of depression recognition in comparison to prior research efforts. This underscores the potential of combining multiple biomarkers to improve the accuracy of depression diagnosis and understanding.

The study of depression through VHs is an area that has been explored in previous research. For example, Philip et al. [278] conducted a study with individuals at various levels of depression, where participants engaged in two randomly ordered interviews, one with a specialist and another with embodied conversational agents (ECAs). In this case, patients with depression exhibited a high degree of acceptance toward the ECAs. User responses were also used to classify the different levels of depression, achieving a sensitivity of 73% and specificity of 95%. This study demonstrates the validity and acceptability of using an ECA for diagnosing major depressive disorders. Egede et al. [279] divided participants into two groups. Half of the participants completed tasks guided by a VH, while the other half performed tasks guided by text on a screen. Various biosignals were used to classify the level of depression in subjects based on the PHQ-9 questionnaire. The results indicated that the use of VHs influenced the expressive behavior of the subjects and improved their disposition towards tasks. Lastly, the study conducted by Takemoto et al. [280] involved a group of participants divided into depressive and non-depressive individuals. They were asked to engage in conversations with both, a VH and human interviewers, with gaze patterns recorded by an eye-tracking device during both types of interactions. The results of the experiment revealed significant differences in eye movements between the control group and the group with depression symptoms.

However, a common limitation across these studies is the absence of a generative language model, such as the current GPT-3 and GPT-4. Moreover, these studies do not employ voice transcription models based on AI, allowing real-time voice conversations between the user and the VH. Nor do they use AI-based voice generation models that can achieve a much more natural and characteristic voice. Finally, they also do not utilize VR platforms that can add greater realism to VHs through aspects like lip or body movement. All in all, these limitations hinder VHs from engaging in spontaneous conversations with subjects, thereby reducing the realism of the interactions and the possibility of eliciting emotions.

In summary, this research focuses on ER and depression recognition through a conversational VHs in VR. It marks the first instance of deploying such an experimental setup for AfC research. It is also the first work that use VHs in the research of mental illnesses such as depression. The study employs various signal processing techniques sourced from ET and EDA data to identify different emotions based on the arousal and valence scale. Additionally, the dataset is enriched by incorporating simple processing of the conversations with the VHs and demographic features. A data classification pipeline, which encompasses statistical methods like feature selection and ML, will be utilized to create the most accurate classification models. To ensure the robustness of the results and prevent overfitting, a validation check has been carried out. Lastly, as the experimentation involves conversations with four different emotional VHs, the same ML pipeline will be employed to determine which emotional VH is more influential in assessing specific targets. This comprehensive approach aims to contribute to the understanding of emotions and mental health assessment within VE.

This manuscript is structured as follows: Section 2 explains the experimental design and the tools utilized therein, along with the analytical methods employed. Section 3 presents the results derived from the aforementioned analysis. Section 4 discusses the work accomplished and the objectives met, and provides a discussion of the results. Lastly, Section 5 concludes the study and highlights the key findings.

5.2 Materials and methods

5.2.1 Participants

A total of 98 participants were recruited for this experiment, with a mean age of 33.95 ± 11.10 years. Out of the participants, 50 were women, 47 were males and 1 identified as another gender. Before their participation, participants were provided with detailed information about the study and provided written consent for their involvement. To protect the privacy of their responses, the data were anonymized and randomized.

Ethical approval for the study was obtained from the Ethical Committee of the Polytechnic University of Valencia (Approval ID: P06_25_07_2022). Data collection took place between October 2022 and February 2023.

To ensure that the participant group was homogeneous and in a healthy mental state, they were assessed using the PHQ-9, a standard psychometric test for measuring levels of depression. The control group, consisting of participants with a PHQ-9 score below 9, included 64 individuals. The mean age of this control group was 31.96 ± 10.34 years. Among them, there were 32 women, 31 males and 1 participant with another gender identity.

5.2.2 Virtual-Human

The VH model utilized in this study is based on the one used in the work of Llanes et al. [272]. In this research, six different VHs were employed, with two of them exhibiting a neutral emotional state, while the remaining four VHs expressed specific emotions, including anger, happiness, sadness, and relaxation. For the purposes of this study, only the data collected from conversations with these four emotionally expressive VHs are considered.

5.2.3 Questionnaire

The experiment began with an initial phase where participants completed a battery of psychological questionnaires that included demographic information, the PHQ-9, state-trait anxiety inventory (STAI) questionnaire, and several other physiological assessment questionnaires, which will be used for further analysis. The STAI questionnaire was utilized to evaluate the emotional baseline of each participant. Afterward, the participants engaged in six different conversations with a set of six VHs. Following each conversation, the participants were instructed to assess the naturalness and realism of the social interaction. Naturalness was evaluated based on the flow of the conversation, ranging from "I felt very forced during the conversation" to "I felt very natural during the conversation." Realism assessed the content generated during the conversation, with ratings ranging from "It was an artificial conversation. It doesn't resemble a real conversation at all" to "It was a realistic conversation. The content was very similar to that of a real conversation." Both metrics utilized a 7-point Likert scale.

Furthermore, participants were required to identify the emotions of both, the VH and themselves by using the self-assessment Manikin (SAM) [237, 238] questionnaire once the conversation with the VH finished. The SAM questionnaire is a standardized measure that assesses valence and arousal, referencing Russell's circumplex model. In this context, participants rated the valence and arousal of both themselves and the VH using two 9-point Likert scales during their conversations. We hypothesized that a happy VH would elicit high arousal and positive valence, a relaxed VH would result in low arousal and positive valence, a sad VH would induce low arousal and negative valence, and an angry VH would lead to high arousal and negative valence.

On the other hand, the patient health questionnaire (PHQ), specifically the PHQ-9, is a widely used psychological test designed to assess and measure the severity of depressive symptoms in individuals [281]. Developed as a self-administered tool, the PHQ-9 is derived from the primary care evaluation of mental disorders diagnostic instrument for common mental disorders. This questionnaire comprises nine questions corresponding to the diagnostic criteria for major depressive disorder. Respondents rate the different questions with a numerical frequency of specific depressive symptoms over the past two weeks, providing a quantitative measure of the individual's depressive state.

The results from the questionnaires in this study were numerical values on different scales. A threshold value was established, with different value depending on the questionnaire, to treat the output of the questionnaires as binary classification problems.

Below each threshold, the score was considered low, and above it, the score was considered high, achieving a binary segmentation. For the questionnaires related to VH naturalness and realism, the threshold was set at 4. Any score exceeding this threshold was deemed high, labeled as 1, and 0 otherwise. The proportions of ones in each target were 66.53% and 62.90% for naturalness and realism, respectively. In the case of arousal and valence scores for subject and VH, the threshold for classification was set to 5, performing the same assignment for the labels as before. The proportion of ones in each target was human arousal 55.24%, human valence 38.31%, VH arousal 58.06% and VH valence 40.32%. For the PHQ-9 questionnaire, specific thresholds were already established for identifying depressive symptoms. A threshold value of 9 was used in the PHQ-9 questionnaire as the point at which a subject can be classified as depressive. If the score was less than 9, the subject was classified as non-depressive (label 0); otherwise they were classified as depressive (label 1). The percentage of depressive subjects was 35.42%.

5.2.4 Data pre-processing

This study encompasses three different types of data sources: ET, EDA, and conversations between humans and VHs, in addition to the demographic variables of each subject. It is explain below how the various data sources have been processed to obtain a unified dataset that enables the performance of different classification tasks.

5.2.4.1 Conversations

Various features have been extracted from the conversation with the VH. These features included the total duration of the conversation, the average time the subject and the VH spend talking, the total number of words spoken by the subject and the VH, count of interactions between the subject and the VH, the maximum number of words and maximum time talking in a single interaction by the subject and the VH. Additionally, other variables were obtained, such as the speech time of the first and last message from the subject in the conversation and the total time spent speaking in the first three and last three messages with the VH. These features collectively provide a comprehensive overview of the conversational dynamics during the interactions between the subjects and the VH.

5.2.4.2 Eye-tracking

The ET signal is processed using the PupilLab algorithm. This algorithm includes several key components for analyzing eye movements and related measurements. The algorithm identifies fixations by considering the eye movement's velocity, where fixations have velocities below $68/s$ and durations exceeding $60\,ms$. It compensates for vestibulo-ocular eye movements using optic flow from the scene camera and applies specific thresholds to distinguish fixations. The algorithm is designed to be robust against head movements and provides information about the total number of fixations and their durations. Eye saccades, which are rapid eye movements between fixations, are characterized by their amplitude in degrees and time duration. Moreover, PupilLab includes a blink detector that identifies blinks by analyzing rapid drops and increases in 2D pupil confidence. This process allows for the computation of the number of blinks and their durations. The system collected data from the gyroscope, providing angular velocity information in the x (horizontal), y (vertical), and z (optical) axes. Additionally, data from the accelerometer was obtained, measuring linear acceleration in the same three axes. Roll and pitch, which describe the device's rotation, were also measured. These processing steps collectively offer detailed insights into eye movements, blinks, and the device's orientation.

Pupil Cloud facilitates automatic gaze mapping for our research by employing the reference image mapper enrichment. This process is grounded in the structure from motion technique, which constructs a 3D model of the environment seen in the scene video and tracks the camera's position within this model. With this information, gaze data is monitored in 3D and then projected onto the reference image plane. To support our analysis, a distinct enrichment was generated for each VH, with tasks being segmented and assigned accordingly. The fixation computation provided positions on static images. AoIs for specific body parts, such as the head, eyes, nose, mouth, left arm, right arm, torso, legs, and feet were defined. Various metrics for each AoI were calculated, including the total fixation count (the sum of individual fixation counts), total fixation time (the sum of fixation durations), time to first fixation (from the conversation's start to the first fixation within the AoI, or the conversation's total duration if no fixation occurs), and total visits to the AoI (the number of consecutive fixation groups within the AoI). These processes enabled to collect comprehensive data on gaze behavior, particularly related to specific body parts, facilitating a detailed analysis of user interactions with the VH.

5.2.4.3 Electrodermal Activity

In EDA signal processing, the automatic recognition and correction model, introduced in [282] to eliminate artifacts in the raw signal, was applied. Subsequently, EDA signal was decomposed into its phasic and tonic components using the *Ledapy* library. This allowed to analyze the corrected EDA signal from three distinct components: the signal itself, phasic and tonic components.

Each EDA source yielded different features. For the corrected signal, statistical attributes including the mean, standard deviation, signal amplitude, and Shannon entropy were computed. Following this, the signal was divided into temporal segments that corresponds to the interactions between human and the VHs. Across each of these EDA segments, the mean, standard deviation and signal amplitude were extracted. Additionally, autoregressive (AR) features with a lag of 4 from the corrected EDA signal were calculated. The number of peaks and the mean value of the peaks was also determined in the phasic component.

Autoregressive features

An AR model is commonly used in fields as statistics or signal processing. The AR model describes the output of the variable as a linear combination of the own previous values of the signal. AR model is specified by p which set the order of the model. An AR(p) is defined in equation 5.1.

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} \tag{5.1}$$

where $\phi_1, ..., \phi_p$ are the coefficients of the AR(p) model. A polynomial degree of $p = 4$ was selected, resulting in the calculation of four coefficients along with the error associated with the AR polynomial fitting. These coefficients were the ones used for the modeling of the EDA segments, using them as features for the ML models.

Wavelet features

The wavelet features were obtained through the transformation of the EDA signal using a discrete wavelet transform (DWT). The DWT of a signal y is computed by passing it through a series of filters. Then the result can be shown as a convolution between the signal and a series of filters showed in equation 5.2.

$$y_{DWT}(i) = \sum_{k=-\infty}^{\infty} y(k)g(i-k) \tag{5.2}$$

where g is the specific filter for the DWT and y_{DWT} id the DWT of the input signal y. For the processing of EDA signal, the Haar window was used, to obtain the wavelet features set.

Hjorth features

Hjorth parameters are indicators of statistical properties used in signal processing in the time domain introduced by Hjorth in 1970 [283]. The parameters are *activity*, *mobility*, and *complexity*. This signal is commonly used in EEG field, however, other experimentations had used it as complementary features for their studies. The Hjorth activity is defined in equation 5.3.

$$activity = var(y(t)) \tag{5.3}$$

where var indicates the variance of the signal in the time domain.

The Hjorth mobility is showed by equation 5.4.

$$mobility = \sqrt{\frac{var\left(\frac{dy(t)}{dt}\right)}{y(t)}} \tag{5.4}$$

Finally, the last feature is the Hjorth complexity in equation 5.5.

$$complexity = \frac{mobility\left(\frac{dy(t)}{dt}\right)}{mobility(y(t))} \tag{5.5}$$

This set of features were extracted and also used as inputs for the ML model.

5.2.5 Data source linkage

The extraction of features was also context-dependent and was based on three distinct conversational situations within the overall experimentation: interactions with different emotional VHs and the time when the human or the VH were speaking. The connection between these diverse data sources is established using the time variable. To begin, the target events were associated with timestamps that indicate when each event occurred. These timestamps were then matched with the timestamps from the other data sources, and the sample with the closest timestamp (in absolute value) to the target event was selected for analysis. Equation 5.6 outlines this methodology mathematically.

$$C_{idx} = argmin(|t_{target} - \vec{t}_{D_x}|) \tag{5.6}$$

where D_x represents the data source to link, t_{target} signifies the target time to establish a connection with, $argmin$ designates the position in the database D_x of the minimum value within the temporal vector, and C_{idx} denotes the nearest position in D_x to the target time. Through this data processing, we can effectively identify the specific data point that corresponds to the occurrence of a particular event, denoted as $target$, within the data source D_x.

In this research, the conversation database was used as a central component, integrating it with the other databases. This approach was chosen because the conversation data contains the essential information about the start and end times of each interaction involving the different emotional VHs. Additionally, it included timestamps indicating when a human subject or a VH was speaking. Therefore, the start and end times of conversations were used as reference points for data extraction from various sources. In cases where temporal records of ET or EDA were missing within the specified time intervals, these gaps were marked as missing data in the final database. Variables with missing data exceeding 10% were excluded from the analysis.

All the previously mentioned features were obtained for each subject and also for each emotional VH. Four data subsets were extracted for each subject because the subject interacted with four emotional VHs. Demographic and questionnaire variables of the subjects were added. Since each subject has four data subsets, the demographic and questionnaire variables, which do not depend on the emotional VH, were duplicated across the data subsets. All in all, the whole database consisted in 392 records where the key columns were the subject and the emotion of the VH.

Table 5.1 presents the number of features obtained for three different scenarios. In the general case, considering the complete data source in the experimentation. Divided by VH, which results in different features for each VH, processing data from the beginning to the end of each emotional VH. Finally, it shows the number of features obtained for each interaction with the VH. Regardless of these three scenarios, various calculations were performed on the variables. A set of information was derived from each variable, including the mean, standard deviation, median, and maximum of that variable. Table 5.1 illustrates the different variable groups in which variables were organized according to each data source and the situation in which the variable was calculated.

In addition, the emotional state of the VH was encoded in just two features, representing high or low valence and high or low arousal. All of these results summarize a final dataset of 392 records with 156 variables for the classification of depressive patients. The dataset for ER had 149 variables, as the information from other questionnaires is

Table 5.1: Description of the set of features obtained by data source. The parenthesis of each cell specifies the total number of features obtained.

Model Variables	**Data Source**	**General**	**by VH**	**by Interaction**	**Total features**
Input	Conversations	-	Num. sentences Time talking Num. diff. words (10)	Num. words Time talking (5)	15
Input	ET	-	Fixations Saccades Blinks AoI Ang. speed (102)	-	102
Input	EDA	Baseline statistics (2)	Statistics AR Hjorth Wavelet (18)	Statistics (6)	26
Input	Demographic	Age Gender Video-games Medicine (6)	-	-	6
Output	Questionnaire	-	Naturalness, Realism, SAM Scores and PHQ-9 (7)	-	7

not utilized to identify the classification of one of them and only 256 rows (64 subjects) due to the elimination of the depressive subjects. The removal of the depressive subjects is because they could bias the ER task due to their emotional state.

One subject was excluded from the analyses of naturalness, realism, emotion elicitation, and emotion identification because their score on the STAI questionnaire exceeded the 95$^{\text{th}}$ percentile (47 for women and 40-41 for men) [239].

5.2.6 Statistical and Machine Learning analysis

5.2.6.1 Statistical and ML tools

Various statistical techniques and ML approaches were employed in this study. To reduce the initial number of features in the dataset, a correlation analysis was conducted. Features with correlations exceeding 0.95 with other features were eliminated, resulting 21 features removed, 14 from ET and 7 from EDA.

Due to the extensive feature set obtained, a feature selection process was employed. For each specific target variable, either a Mann-Whitney test or an unpaired t-test was applied to determine which features held statistical significance concerning the target. Multiple confidence intervals were established based on common p-value thresholds, including 0.05, 0.01, and 0.001, resulting in various sets of features depending on the chosen statistical significance level. In cases where no target variable displayed significance with a p-value below 0.05, an alternative feature selection method was employed. This method involved utilizing mutual information coefficient (MIC) [284] when the target variable was discrete. MIC measures the interdependence between two distinct variables, producing a value of zero when the variables are entirely independent and a higher value when there is greater dependence between them. Specifically, the first 30 features with higher values and above zero were chosen as significant variables based on mutual information score. If none of these selection methods yielded a single significant variable, the complete set of features was retained.

Different ML algorithms were employed to explore the optimal selection of features, models and parameters for classifying whether the subject had a high or low score in the target variables under investigation. The models used were: SVM [285] with parameters C (0.01, 0.1, 1, 10, 100), γ (0.01, 0.1, 1, 10, 100), and kernels(*rbf*, *sigmoid*); Logistic Regression (LogR) with parameter C (0.01, 0.1, 1, 10, 100); RFC [286] with parameters for the number of estimators (100, 200, 400) and maximum depth of (5, 10, 20, 30, or without a maximum depth).

5.2.6.2 ML pipeline

The ML pipeline designed for ER and depression recognition consists in a two nested cross-validation (CV) splits. The inner and outer CV had 10 folds and 4 repetitions. Evaluation metrics, such as accuracy, kappa, precision and ROC-AUC were scrutinized on the test set following the model obtained from the inner partition. These metrics were computed for 40 different combinations of test sets, yielding mean and standard deviation values for each combination. The inner combination used unique sets of features and models for each train-validation iteration, followed by a grid search which model parameters were optimized by the balanced accuracy score. The model was then fitted with the training data and applied to the test set. Figure 5.1 provides a visual representation of the described processing pipeline.

In the context of naturalness, realism and ER, each subject exhibits distinct emotional responses based on the VH entity with which they engage. Consequently, individual subjects possess varied targets contingent upon the emotional state of the VH.

On the other hand, the case study of depression recognition was not conducted on a sample-by-sample basis. A single subject may correspond to multiple samples due to interactions with one or more VHs (after data processing, several samples belonging to the same subject may be removed) but it always had the same target. Therefore, the partitioning into training, validation, and test sets was based on subjects and not by individual samples. From the ML pipeline described earlier, the two CV were replaced by group cross validation (GCV), wherein samples were grouped based on each subject. To obtain the prediction in the ML pipeline, the probabilities of being depressive for each of the subject's samples were calculated. The mean probability across all samples was computed. If this mean exceeds 0.5, the subject is classified as depressive; otherwise, they were classified as non-depressive.

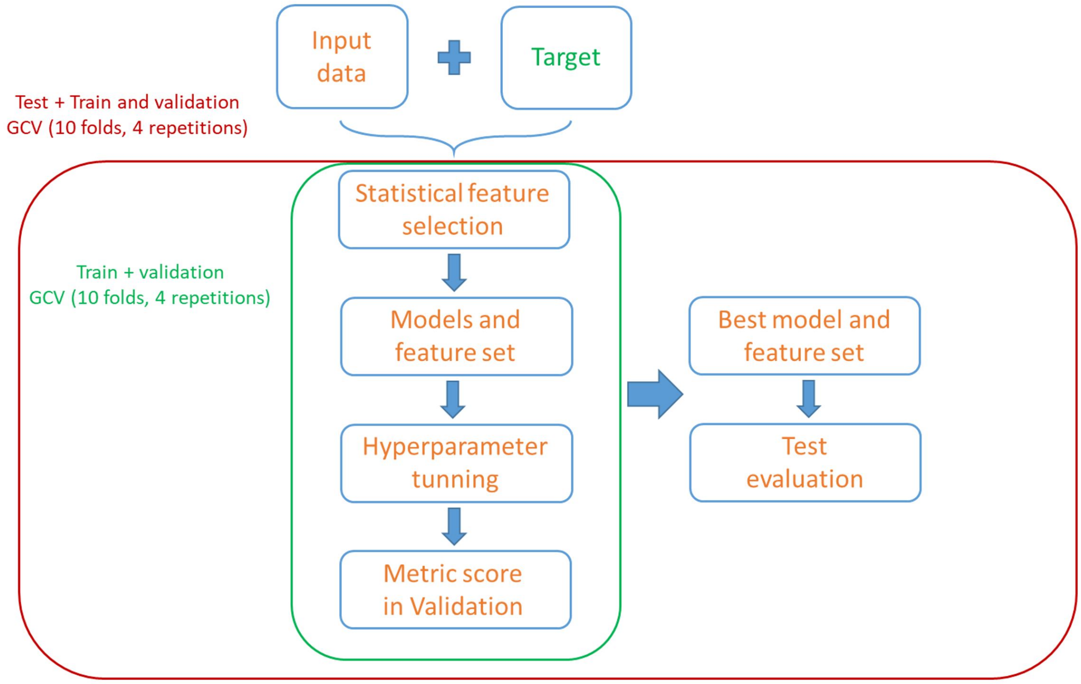

Input data
Target
Test + Train and validation
GCV (10 folds, 4 repetitions)
Train + validation
GCV (10 folds, 4 repetitions)
Statistical feature selection
Models and feature set
Hyperparameter tunning
Metric score in Validation
Best model and feature set
Test evaluation

Figure 5.1: Scheme of the ML pipeline used for social ER and depress recognition.

The same pipeline is employed to build a ML model for each emotional state (Angry, Happy, Sad, and Relax) for each corresponding target. This division aims to determine which emotional state of the VH yields the highest classification metrics. Additionally, the ROC curves of the four different emotional VHs were computed, adding also the ROC curve of the general model that simultaneously considers all four distinct emotional states.

5.2.6.3 Overfitting check

To assess the potential overfitting of the ML pipeline, six random targets were created. The ML pipeline was trained on each of these random targets. These random targets were designed to mirror the proportion of ones found in the actual target. The goal of these methodology is to compare statistically the score obtained by the random targets against the real target. If the comparison between these distributions reveals a statistically significant difference (indicated by a *p*-value of less than 0.05), it suggests that the ML pipeline is performing beyond a chance level and the results are not overfitted. An unpaired two samples t-test was used as statistical test.

5.3 Results

5.3.1 Naturalness and Realism

The results for naturalness and realism are presented in Table 5.2. This table displays the outcomes for the different target variables, indicating the combination of model and feature subset that achieved the highest balanced accuracy in each target classification. The highest balanced accuracy is achieved for the realism target with 0.717. This target also attains the best results for the other scores, except for precision and TNR. This table also contains the overfitting score, showing that both targets have all the metrics with high statistical difference against the random targets, except the TNR metric. The selected models are SVC and RFC. The feature selection method for both targets is All.

Table 5.2: Results of the models obtained with highest balanced accuracy for the different ER targets. These results are depicted in terms of the mean and standard deviation, calculated over the test set for each individual subject. The overfitting check is showed through the significance level with the score. The statistical results, between parenthesis, are shown as - no significant difference, * *p*-value < 0.05, ** *p*-value < 0.01 and *** *p*-value < 0.001.

Target	Feature selection	Model	Balanced accuracy	Kappa	Precision	ROC-AUC	TPR	TNR
Naturalness	All	SVC	0.697 ± 0.105 (***)	0.364 ± 0.192 (***)	0.810 ± 0.117 (***)	0.697 ± 0.105 (***)	0.770 ± 0.141 (***)	0.624 ± 0.234 (-)
Realism	All	RFC	0.717 ± 0.094 (***)	0.419 ± 0.171 (***)	0.777 ± 0.118 (***)	0.717 ± 0.094 (***)	0.831 ± 0.098 (***)	0.602 ± 0.193 (-)

The results obtained for each emotional VH are presented in Table 5.3. The table displays the results for each target and metric, categorized by emotional state. The last column of the table reveals the statistical results in terms on the *p*-value, obtained from the comparison of each metric distribution for the different emotional states. For naturalness the best VH is the anger VH in terms of balanced accuracy and kappa. On the other hand, for realism, the best VH is the happy one in terms of kappa, achieving a 0.175, but the anger is the best model according with the balanced accuracy with a 0.655.

Table 5.3: Results of the emotional VH models with highest balanced accuracy. Different scores are shown for each VH emotion. Results are shown as mean and standard deviation over test set per subject.

Target	Metric	Emotional state			
		Anger	Happy	Sad	Relax
Naturalness	Balanced Accuracy	0.741 ± 0.216	0.641 ± 0.24	0.643 ± 0.226	0.581 ± 0.309
	Kappa	0.229 ± 0.353	0.237 ± 0.393	0.091 ± 0.288	−0.051 ± 0.322
	Precision	0.702 ± 0.313	0.724 ± 0.266	0.726 ± 0.264	0.525 ± 0.406
	ROC-AUC	0.646 ± 0.19	0.664 ± 0.211	0.55 ± 0.15	0.473 ± 0.196
	TPR	0.936 ± 0.154	0.858 ± 0.222	0.864 ± 0.308	0.393 ± 0.375
	TNR	0.359 ± 0.418	0.401 ± 0.414	0.2 ± 0.4	0.517 ± 0.452
Realism	Balanced Accuracy	0.655 ± 0.266	0.575 ± 0.354	0.594 ± 0.195	0.576 ± 0.356
	Kappa	0.141 ± 0.442	0.175 ± 0.582	0.071 ± 0.258	0.154 ± 0.574
	Precision	0.635 ± 0.377	0.730 ± 0.395	0.772 ± 0.207	0.514 ± 0.416
	ROC-AUC	0.596 ± 0.251	0.650 ± 0.374	0.542 ± 0.138	0.594 ± 0.352
	TPR	0.778 ± 0.302	0.574 ± 0.369	0.929 ± 0.175	0.812 ± 0.348
	TNR	0.389 ± 0.454	0.567 ± 0.389	0.083 ± 0.276	0.379 ± 0.439

Appendix B complements the information showed in this section, showing the distribution plots of the various metrics using boxplots. It is also included the ROC curves for each emotional state, along with the general model.

5.3.2 Social emotion recognition

The results for the various social emotions are presented in Table 5.4. This table displays the outcomes for the different target variables, indicating the combination of model and feature subset that achieved the highest balanced accuracy score in each target classification. The highest balanced accuracy obtained is achieved for the VH valence with a 0.655 score. This target also attains the best results for the other scores, except for precision, TPR and TNR. The lowest balanced accuracy and kappa is obtained for human valence target. The overfitting test show significant differences specially in VH arousal target. Conversely, between two and three significant differences could be obtained in the analysis of the other targets. The models selected were RFC and SVC two times both of them. The feature selection was also selected two times MIC and two times all the variables.

Table 5.4: Results of the models obtained with highest balanced accuracy for the different ER targets. These results are depicted in terms of the mean and standard deviation, calculated over the test set for each individual subject. The overfitting check is showed through the significance level with the score. The statistical results, between parenthesis, are shown as - no significant difference, * p-value < 0.05, ** p-value < 0.01 and *** p-value < 0.001.

Target	Feature selection	Model	Balanced accuracy	Kappa	Precision	ROC-AUC	TPR	TNR
Human Arousal	MIC	RFC	0.581 ± 0.128 (-)	0.153 ± 0.247 (-)	0.593 ± 0.145 (**)	0.581 ± 0.128 (-)	0.696 ± 0.155 (***)	0.467 ± 0.175 (***)
Human Valence	All	RFC	0.553 ± 0.124 (-)	0.106 ± 0.262 (-)	0.488 ± 0.327 (*)	0.553 ± 0.124 (-)	0.305 ± 0.199 (-)	0.802 ± 0.133 (-)
VH Arousal	MIC	SVC	0.630 ± 0.112 (**)	0.234 ± 0.210 (**)	0.675 ± 0.148 (***)	0.630 ± 0.112 (*)	0.671 ± 0.174 (**)	0.589 ± 0.217 (-)
VH Valence	All	SVC	0.655 ± 0.082 (-)	0.292 ± 0.158 (-)	0.536 ± 0.111 (*)	0.655 ± 0.082 (-)	0.644 ± 0.130 (***)	0.666 ± 0.072 (***)

The results obtained for each emotional state of the VH are presented in Table 5.5. The table displays the results for each target and metric, categorized by emotional state. Thre results show that relax and sad VH obtains most of the times the highest scores in terms on Kappa and ROC-AUC scores. The relax VH obtains the best results for the identification of human arousal and valence with a kappa score of 0.288 and 0.171 respectively. Whereas, sad VH obtains the best results for VH arousal and valence with a kappa score of 0.5 and 0.178 respectively. Indeed the relax VH model surpasses the scores of the base model which contains the information from the rest of the emotional VH.

Furthermore, Appendix B provides several visual representation of the various metrics and targets distributions using boxplots for the division by emotional states and also, the distributions obtained for the overfitting test. The Appendix also includes the ROC curves for each emotional state, along with the general model.

5.3.3 Depression recognition

Table 5.6 shows the results for depression recognition for the best model. The results are averaged over the test set of the different folds. Is also shown the best set of features obtained for the model, according the feature selection method implemented. The best model is the SVC with the features from the 0.05 ($*$) significance level. The balanced accuracy obtained exceeds the 0.65 score, which is 0.685, while kappa score is 0.388 and

Table 5.5: Results of the emotional VH models with highest balanced accuracy. Different scores are shown for each VH emotion. Results are shown as mean and standard deviation over test set per subject. Statistical comparison between the different distribution is shown as well.

Target	Metric	Emotional state			
		Anger	Happy	Sad	Relax
Human Arousal	Balanced accuracy	0.611 ± 0.314	0.732 ± 0.236	0.561 ± 0.238	0.635 ± 0.267
	Kappa	0.067 ± 0.249	0.046 ± 0.079	−0.006 ± 0.160	0.288 ± 0.410
	Precision	0.095 ± 0.233	0.571 ± 0.495	0.167 ± 0.338	0.642 ± 0.334
	ROC-AUC	0.550 ± 0.150	0.531 ± 0.054	0.500 ± 0.094	0.669 ± 0.214
	TPR	0.333 ± 0.471	0.062 ± 0.108	0.133 ± 0.287	0.866 ± 0.282
	TNR	0.641 ± 0.423	1.0 ± 0.0	0.844 ± 0.341	0.417 ± 0.479
Human Valence	Balanced accuracy	0.485 ± 0.290	0.520 ± 0.167	0.510 ± 0.309	0.680 ± 0.279
	Kappa	0.135 ± 0.364	0.049 ± 0.263	0.0 ± 0.0	0.171 ± 0.353
	Precision	0.353 ± 0.361	0.188 ± 0.370	0.021 ± 0.081	0.132 ± 0.318
	ROC-AUC	0.604 ± 0.216	0.531 ± 0.145	0.500 ± 0.000	0.604 ± 0.190
	TPR	0.567 ± 0.442	0.144 ± 0.272	0.111 ± 0.314	0.25 ± 0.433
	TNR	0.500 ± 0.423	0.882 ± 0.191	0.667 ± 0.453	0.869 ± 0.283
VH Arousal	Balanced accuracy	0.625 ± 0.346	0.724 ± 0.252	0.554 ± 0.226	0.768 ± 0.221
	Kappa	0.093 ± 0.250	0.059 ± 0.235	0.050 ± 0.163	0.080 ± 0.544
	Precision	0.25 ± 0.417	0.822 ± 0.193	0.213 ± 0.337	0.718 ± 0.399
	ROC-AUC	0.604 ± 0.259	0.518 ± 0.12	0.523 ± 0.177	0.762 ± 0.212
	TPR	0.286 ± 0.41	0.973 ± 0.083	0.346 ± 0.455	0.788 ± 0.286
	TNR	0.875 ± 0.298	0.053 ± 0.223	0.642 ± 0.395	0.654 ± 0.367
VH Valence	Balanced accuracy	0.535 ± 0.269	0.573 ± 0.245	0.612 ± 0.152	0.825 ± 0.228
	Kappa	0.017 ± 0.452	0.020 ± 0.353	0.178 ± 0.245	0.050 ± 0.150
	Precision	0.667 ± 0.315	0.255 ± 0.361	0.534 ± 0.375	0.040 ± 0.116
	ROC-AUC	0.500 ± 0.274	0.513 ± 0.217	0.617 ± 0.153	0.537 ± 0.105
	TPR	0.759 ± 0.339	0.373 ± 0.430	0.500 ± 0.341	0.333 ± 0.471
	TNR	0.200 ± 0.400	0.652 ± 0.360	0.724 ± 0.325	0.747 ± 0.388

the ROC-AUC score is above 0.65 with a final score of 0.685. Finally, the TPR and TNR scores are 0.542 and 0.828 respectively.

The results of depression recognition splitted by the VH emotional state are shown in table 5.7. The scores obtained shown that the closest models to the base model are happy, relax and sad VHs, being the relax VH the one which obtains the best results in terms of balanced accuracy, kappa, ROC-AUC and TNR. The anger model appears to be the model with lowest scores.

Figure 5.2 shows a set of boxplots for the different computed metrics. The four different emotional states of the VH have similar distributions. However the sad state

Table 5.6: Result of the best model for depression recognition. Mean and standard deviation are shown for each score over the test set per subject. Added to the numeric results is shown the level of significance of the comparison between each metric and the overfitting check inside parenthesis. The statistical results are shown as - no significant difference, * p-value < 0.05, ** p-value < 0.01 and *** p-value < 0.001.

Model	Feature set	Balanced accuracy	Kappa	Precision	ROC-AUC	TPR	TNR
SVC	0.05	0.685 ± 0.199 (**)	0.388 ± 0.413 (***)	0.733 ± 0.351 (***)	0.685 ± 0.199 (***)	0.542 ± 0.281 (***)	0.828 ± 0.239 (-)

Table 5.7: Results of the different emotional VH models with highest balanced accuracy obtained. Different scores are shown for each VH emotion. Results are shown as mean and standard deviation over test set per subject. Statistical comparison between the different distributions obtained by each emotion is shown. The statistical results are shown as - no significant difference, * p-value < 0.05, ** p-value < 0.01 and *** p-value < 0.001.

Metric	Emotional state			
	Anger	Happy	Relax	Sad
Balanced accuracy	0.584 ± 0.187	0.637 ± 0.128	0.646 ± 0.143	0.613 ± 0.145
Kappa	0.182 ± 0.365	0.252 ± 0.242	0.254 ± 0.286	0.226 ± 0.274
Precision	0.533 ± 0.31	0.548 ± 0.328	0.495 ± 0.413	0.582 ± 0.242
ROC-AUC	0.584 ± 0.187	0.637 ± 0.128	0.646 ± 0.143	0.613 ± 0.145
TPR	0.454 ± 0.239	0.583 ± 0.327	0.517 ± 0.425	0.65 ± 0.241
TNR	0.713 ± 0.256	0.69 ± 0.311	0.775 ± 0.281	0.577 ± 0.322

shows the most compact distribution. However the boxplots does not show clearly any emotional state which distributions are, consistently, above the rest of states.

Figure 5.3 depicts the AUC curve of the various models studied for depression recognition. In this case, the four different emotional VHs have similar trajectories. In contrast, the general model is above all of them achieving a higher score.

The comparison against random targets shows statistical differences between 0.05 and 0.001 in metrics such as kappa, precision, ROC-AUC and TPR. The distributions of the metrics are shown as boxplots for the real and the random targets in Figure 5.4. In general, the performance of the model with random targets tended to be lower than the model trained with the depression target. Furthermore, the metrics, such as accuracy, tended to approach the proportion of ones in the real target and also, were closer to a score of 0.5 in the ROC-AUC score or a low score in TPR, indicating the random behavior of the ML pipeline when random targets are studied.

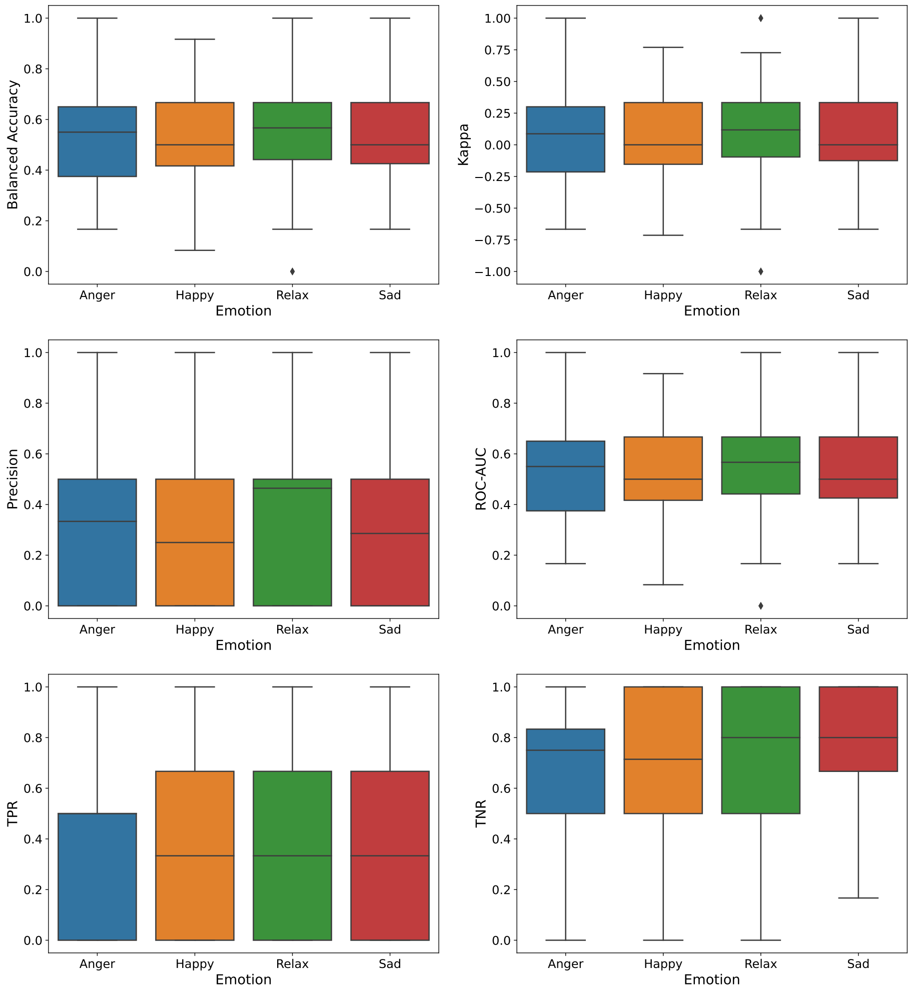


Figure 5.2: Boxplots distribution of the different scores obtained in the test set per subject for the different emotional VH state.

5.4 Discussion

This study represents the first research to evaluate and model ER and depression assessment in real-time conversations with a VH using biomarkers and AI. This experimenta-

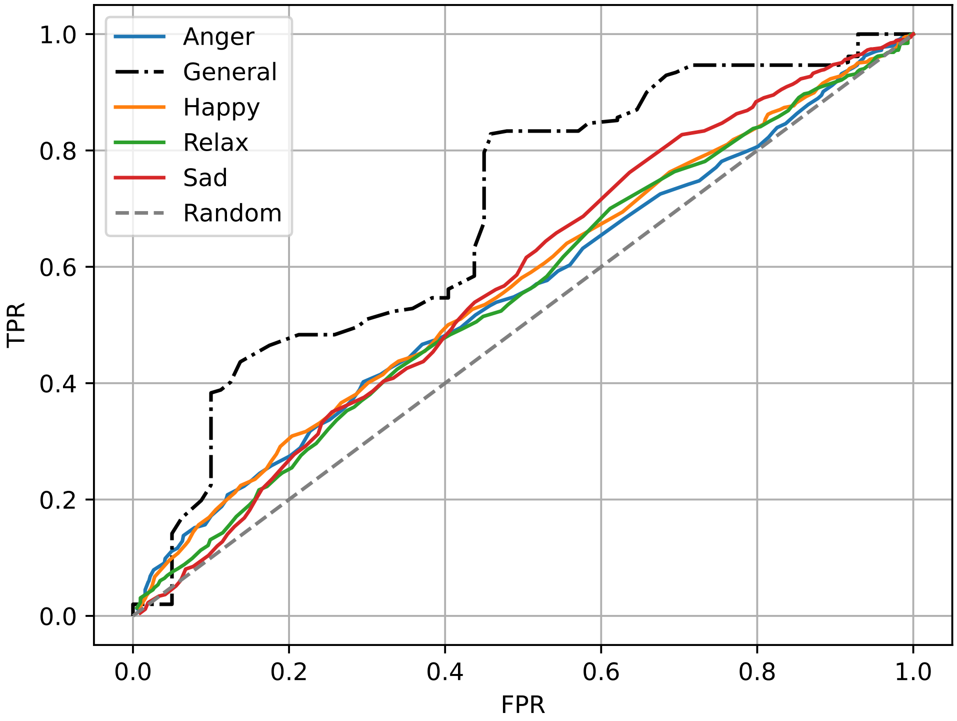


Figure 5.3: AUC curve of the different models in terms of the VH emotional state. The general model is in black dashed line. The diagonal line in grey indicates the performance in the AUC curve of a random model.

tion represents a novel application of an intelligent system that improves the ability to generate social affective processes. Moreover, various signal processing techniques have been applied to maximize the information gleaned from different physiological features, depending on the type of signal. Several variable selection techniques and ML models have been tested to achieve the most accurate model. Across all these tasks, the results highlight the potential of this type of experimentation for identifying subjects emotions effectively and recognizing depressive patients.

5.4.1 Naturalness and Realism

Regarding the results obtained in naturalness and realism, models with high values of balanced accuracy with a 0.697 and 0.711 and kappa of 0.364 and 0.419 are achieved for

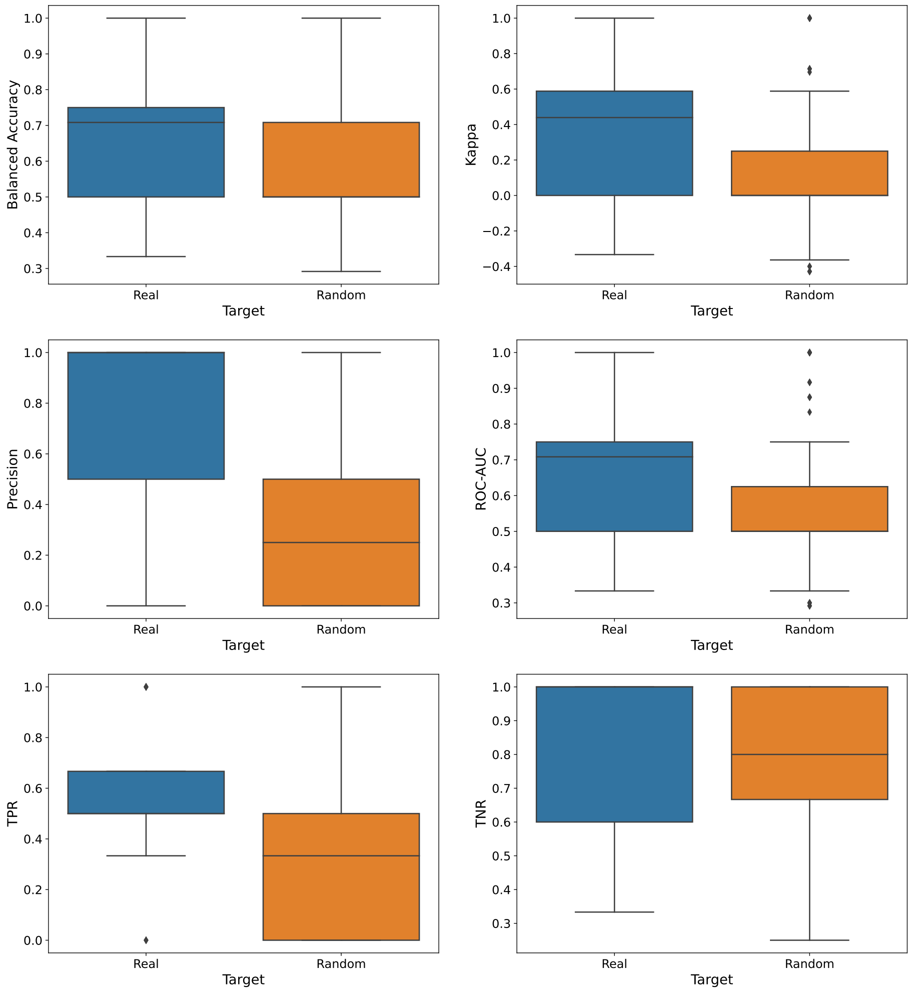


Figure 5.4: Distributions of the different scores obtained for each prediction on the test set. The dark blue represents the model obtained using the real target, while the orange represents the model obtained using a random target.

naturalness and realism respectively. The overfitting check reveals significant differences in all metrics except one. This check demonstrates that the obtained models are not overfitted and show results over the chance level. It can be concluded that the recognition of naturalness and realism carried out, through conversations with VHs, is robust and yields good results.

The VH that achieves the highest values in the metrics is mainly the anger VH with a balanced accuracy of 0.741 and 0.655 for naturalness and realism respectively. This indicates that the anger VH exhibits the highest effectiveness in representing naturalness and realism. This may stem from the possibility that participants in such interactions can attain higher level of empathy in comparison to interactions with other emotional states of the VH. Indeed, in the case of naturalness, the anger model achieves better results that the general model. The second higher score, in general, is achieved by the happy VH in both targets. These two results indicates that emotional states of higher arousal, through the subject's interaction with the VH, could be recognize better and robustly. Therefore, emotional states of greater intensity can generate a greater sense of realism and naturalness, regardless of the valence value of the emotion.

5.4.2 Emotion recognition

The results obtained for ER, they are worse compared to naturalness and realism. In this case, the overfitting check does not turn out to be significant in most metrics. This demonstrates that these models are not as robust as in the previous case of naturalness and realism. However, robust results are achieved for the study of VH arousal and valence, achieving a balanced accuracy of 0.630 and 0.655 respectively. However, the identification concerning emotion elicitation does not turn out to be entirely robust.

Regarding the results obtained for each emotional VH, in the case of arousal and valence elicitation, the results obtained by the relax VH outperform the general model, with a balanced accuracy of 0.635 and 0.680 respectively. This demonstrates that the use of more than one VH may not be beneficial for emotion elicitation. Regarding the identification of the emotional state of the VH, the result is also not entirely robust for any emotional VH. In any case, the sad VH achieves the best results in the case of valence identification with a balanced accuracy of 0.612 and a kappa result of 0.178.

Comparing the work performed against previous researches, this study contributes with several novelties to the state of the art. As has been seen, the works of [269], [270] and [271] used VH to study ER. However, these VHs do not have any interaction with the subject or they make an interaction using a prepared script for a conversation. This methodology differed highly with the one exposed in this work, where the VH, is completely autonomous and could make any kind of answer. Moreover, the emotions displayed by the VHs in these studies was induced through the avatar of the experimentation. Another important novelty of this work is that it makes this implementation but also, the emotional state is also induced through the contextualization of the LLM used for answer generation. And finally, this work is the first that studies the ER through ML models which are based on biosignal features. Past works did not used biosignals to perform ER in this kind of experiments.

In conclusion, the results of the ER task do not demonstrate the desired level of robustness required for strong classification evidence. This limitation may stem from several factors. Firstly, the lack of modulation in the VH's voice according to the emotions conveyed in its sentences could potentially diminish the empathetic response from the subject. Similarly, the context provided for each VH may not have been sufficiently nuanced to foster empathy with the VH. Additionally, it is important to explore alternative signals such as EEG or HRV, which could offer complementary information beyond what was analyzed in this study. Furthermore, a more thorough analysis of the subject's voice and textual content could enhance the assessment of displayed emotions. These data sources may offer the most informative insights to improve the performance of ER tasks.

5.4.3 Depression recognition

The results obtained in depression recognition are indicative and promising as an initial exploration. The recognition of depressive subjects is consistent, achieving a balanced accuracy score of 0.685 and a kappa score of 0.388. Nevertheless, a high TNR which value is 0.828 was achieved. This is significant because the model's ability to assign a subject as non-depressive is very consistent. In other words, if the model categorizes you as non-depressive, there is a high likelihood that you are not. This also implies that it does not exceeding the classification of depressive subjects, making it a reliable model. The overfitting check achieve significant results for all metrics, showing an increment of the results above the chance level. However, the standard deviation of the results shows to be high compared with other models. More tests and research should be performed to corroborate the results obtained in this work. In conclusion, the model tested in this work could therefore provide reliable guidance for future researchers in depression recognition through VH.

On the other hand, the models obtained by emotional VHs achieve results similar to the general model, especially the happy, relax, and sad. The anger VH performs the worst. The highest result is achieved by the relax VH with a balanced accuracy of 0.646. Therefore, the recognition of depressive patients may be easier when it comes to eliciting emotions other than anger, with those having more positive valence showing better results. This could be due because depressive subjects find more difficult to empathy with other subjects which emotional state has high valence.

The methodology followed in this work improves depression recognition experiments performed before. In this work an AI-based VH is developed to maintain real time conversations with subjects. The conversations did not followed any pattern or question script, the VH was under no boundary. The experimentations performed before such as [278], [279] and [280] had a script of questions for the VH, avoiding the freedom of the conversations. Moreover, these studies compared the performance of the interview of the VH with the analogous performed by a specialist. Moreover, the work of Takemoto et al. [280] is the only one that uses ET to analyze the differences between the subjects interviewed by the VH or the specialist. Therefore, this study contributes with several novelties to the state of the art. This is the first study that does not perform an interview over the subjects, on the contrary, this work developed an autonomous VH without any script constrain. Secondly, is the first study that ensembles different biosignals to perform depression recognition in this type of experimentation, contributing with a methodology to study ML depression classification models.

5.4.4 Limitations and future research

To enhance the experimentation for ER or depression recognition, several improvements can be implemented. Firstly, the contextualization of the VHs should improved or modified for certain tasks, depending on the task to be performed. For example, the VH with anger context should be modified for depression recognition, but for ER it is more necessary to modify the context of the happy VH. The introduction of personal questions could be a first step to improve emotion elicitation during the conversations. On the other hand, improvements are needed in emotion elicitation. Testing different text-to-speech models or using different types of messages could enhance emotional elicitation from the VHs. Additionally, the synthesized voice did not modulate rhythm, tone, or volume based on emotions in the message. The VH's body movement was limited to idle motions, significantly reducing the scope of bodily expressions, and facial expressions remained static throughout the task. Overall, improving these factors can lead to greater effectiveness in emotion elicitation. In the case of depression recognition, the obtained results are promising and interesting. However, more research should be perform to verify the obtained results in this work.

The relevance of this experimentation is that it allows a high level of freedom for the subject. Therefore, to model the experimentation correctly, it needs more tools to generalize all the information collected. One important factor is the number of subjects. More subjects are needed to study to generalize patterns of the different type of conversations. A higher number of subjects could generalize better this type of experimentation. Moreover, the experimentation perform allows the subject to express movement, gesture and sentences in many different ways. Fundamental knowledge could be find in features or data sources that are not studied in this work. For this reason, a higher work in feature creation and source analysis should be performed to find more interesting variables to use in the ML models.

From the modelling part, there are few improvements that could be done. The first of them is the study of other different feature selection methods. This work also experimented with feature selection methods based in recursive feature elimination. However, these results eventually led to overfitting and did not achieve a proper performance on the test set, discarding them. It has been observed that simpler variable selection methods, such as checking the correlation of these variables with the target, result in less overfitting of the model performance. In addition, this work did not use any imputation method. The use of imputation methods would increase the data samples, avoiding the exclusion of subjects with only a few missing variables. This would be particularly important in the case of studying models based on the VH emotional state, where the number of samples is reduced, potentially hindering good model generalization.

On the other hand, more in-depth research should be conducted on the features related to conversations with the VHs. In our case, the features extracted from this data source was lower compared with the ET and EDA sources. However, the authors of this study believe that these conversations contain more information than is currently being captured. Further processing, such as sentiment analysis in the sentences or obtaining the most frequent words could be an effective step in giving this data source more relevance. Indeed, a LLM could be use in order to analyze the plain text of the conversation. The LLM could perform sentence recognition or conversation summarizing or also, it could obtain strategic outputs to use them as variable inputs for the ML model.

5.5 Conclusion

The work presented represents a significant advancement in AfC research and in the recognition of depressive subjects through the use of VHs. This study analyzes emotion and depression recognition with the involvement of four different emotional VHs, utilizing features extracted from ET, EDA, and conversational data. The research explores various feature extraction methods and establishes a methodology to identify the optimal set of features and models for each target. The findings in this work reveal certain trends and patterns, particularly in the recognition of depressive patients, which pave the way for utilizing VHs in identifying these targets. The obtained results are promising and further evidence is needed in order to implement procedures that would improve the objectivity and validity of ER and depression assessment. However, the findings suggest that there is certain knowledge from the ET and EDA biomarkers that would allow the automatizing of, especially, depression assessment. The system and methodology employed in this study have a multitude of potential applications in various fields, not limited to healthcare or AfC but extending to areas such as education or psychology.

Chapter 6

Discussion

This chapter discusses the major implications of this thesis. The overall results of the thesis are examined, placing them in the context of the ultimate goal of the thesis. Finally, various future research directions related to the different fields studied are presented.

6.1 General framework

The primary aim of this thesis is to construct an affective computing (AfC) system geared towards eliciting and recognizing emotions through the use of automatically processing biosignals such as eye-tracking (ET) and electrodermal activity (EDA). To elicit emotions, a virtual human (VH) capable of engaging in real-time conversations within a subject is developed in virtual reality (VR). The VH was developed using the latest AI technologies to allow natural VH-human interactions such as large language model (LLM) to generate VH answers, lip synchronization or speech to text model. Finally, these developments were ensembled to perform emotion elicitation and recognition and, additionally, depression recognition.

Firstly, this thesis studies the adaptation of an ET fixation detection algorithm from 2D to 3D in VR. To adapt the algorithm into 3D VR, head movements were taken in consideration to compute the subjects gaze in the virtual environment (VE). Moreover, this work also presents a methodology to calibrate the parameters of the algorithm in a VE. Notably, this calibration approach is algorithm-agnostic, relying instead on ET features like fixation count or mean fixation duration.

However, one objective of this thesis was to utilize the ET algorithm alongside automatic feature extraction. Nevertheless, the algorithm could not be used for automatic fixation identification during the VH experimentation phase. This limitation arose because the VH was developed in a semi-immersive setup due to graphical technical restrictions. Considering this, the algorithm developed for full immersive VR with HMD, could not be applied under these conditions. For this reason, the algorithm was not utilized in the final experiment.

Despite this, thanks to the insights gained in the algorithm and methodology developed, a deeper understanding from the ET field has been obtained in order to understand the features set extracted by PupiLab. Therefore, even though the fixation identification algorithm developed could not be directly applied, it has offered valuable knowledge necessary to identify features and understand which could be relevant for addressing issues related to ER or the classification of depressive subjects.

The primary factor that has influenced the inability to construct the VH in immersive VR is the technological development of VR. Despite the enormous progress made in this field in recent years, and its increasingly commonplace use, there are some technologies that are challenging to develop in this environment. In the case of VH development, it was possible to create a virtual avatar movement allowed, but there was no way to include the realistic and natural lip movement from real-time voice audio. Although Unity has libraries like Salsa that are starting to implement these features, they could not implemented at the date of the experimentation started to be tested. Alternatively, there was Audio2Face from Omniverse. This library allowed the development of a VH that could not only have simple movement but also move the lips in real-time and in sync with the input audio involving highly realistic simulation. In this way, a realistic as possible VH could be created. However, this library is not allowed for immersive VR, being only possible to use in screens. It was decided to proceed with this approach for the experimentation, even if it meant foregoing immersive VR and, consequently, the implemented ET algorithm.

Secondly, the EDA artifact detection and correction algorithm, also represents a breakthrough in the analysis of this signal. The designed algorithm is the first on that uses deep learning (DL) models for the recognition of artifacts and also, it presents an algorithm to correct them. The results are not only innovative for this reason, but they also overcome state of the art results, establishing the comparison between machine and human correction as a metric to measure and optimize. This algorithm was applied to the VH experimentation phase. The EDA correction algorithm could eliminate artifacts from the signal without the need of manual intervention. This tool has significantly reduced the time required for developing the experimentation and processing data, but especially, it enables the automatic processing of EDA signals for use in assessment tasks.

On the other hand, this thesis contributes with another significant result. To the best of our knowledge, this work has not only been the first to construct a VH using the most advanced AI technologies available to date but has also achieved the objective of conducting emotion recognition (ER) and the identification of depressive patients through the use of conversational VHs.

For the construction of the VH, it has been necessary to understand and adapt various recently innovative AI models, such as Nvidia Omniverse, Google S2T, and especially Large Language Models (LLM) like GPT-3. This work assembles all these technologies into different modules, allowing the replacement of the AI model of any of them with a much better or more innovative one. Furthermore, this work demonstrates how to integrate all these modules to design the most natural VH possible. The advances achieved in this work are a first step that opens the door to the design, implementation, and use of VH in VR, something that had not been done before.

Lastly, this thesis successfully utilizes the designed VH and contextualizes it with various emotions and aspects to determine if its use can elicit emotions. A technical analysis of the conversations is performed. The results showed the differences between the sentences and interactions between the human and the VH, while other metrics are computed such as the amount of errors produced by the LLM. In general, most of the conversations are fluid, achieving a high percentage of conversations above the 4 minutes of duration.

This experimentation also allowed a second type of analysis. This other work studied the ER and depressive patterns through the data collected during the conversations, which also involves EDA and ET signals. After the use of different signal processing and feature extraction techniques, a set of variables was obtained to predict, through machine learning (ML) models, certain targets from the ER and a depression target. The results of the analysis showed a good performance in the naturalness and realism identification, but lower scores in arousal and valence recognition. However, the results obtained for depression recognition showed to be promising achieving a high rate of accuracy and precision.

The results provided by this thesis are numerous and encompass various fields, ranging from signal processing and the study of statistical and ML methods to the application in an experiment involving a VH. Moreover, the work presented in this thesis manages to make novel and significant contributions to the different studied fields.

6.2 The use of biosignals for emotion recognition

As it has been explained in this thesis, there are numerous studies that employ various biosignals to conduct ER or pattern recognition related to health, such as stress or, in our case, depression. The use of these biosignals provides an objective understanding of the subject's behavior that cannot be controlled by the subject themselves. This is in contrast to questionnaires, where the subject's response is subjective and may, therefore, complicate the discovery of behavioral patterns. For this reason, the study of biosignals is becoming increasingly important in conducting psychological or health experiments.

In this thesis, two signals have been studied which are ET and EDA. For the study of the ET signal, a methodology based on the use of metrics related to fixations and AoIs has been developed to calibrate an ET fixation detection algorithm. In this case, different parameter values have been investigated to find optimal points that achieve the best possible results in various metrics simultaneously. Different sets of parameters are obtained, with the 1° dispersion and 0.25 s window time parameters being the best set, achieving a percentage of points classified as fixations of 67.82%. However, this is a specific case for the type of VE under study. This work primarily provides a methodology to follow for achieving the identification of fixations and saccades that respects various metrics simultaneously and is algorithm-agnostic.

On the other hand, this work also introduces a DL model that automatically identifies artifacts in the EDA signal and corrects them automatically using a regression algorithm. This is the first work that achieves both objectives, automatic detection and correction of artifacts in EDA signal. This work contributes numerous innovations that result in a final outcome where EDA artifacts are recognized with a 72% TPR and an accuracy of 87%. It also succeeds in surpassing several widely used state-of-the-art models. Subsequently, the correction performed by the regression algorithm on the detected artifacts is analyzed. The work statistically demonstrates that the signal corrected by the regression algorithm is not significantly different from the manually corrected signal, while both types of corrections are different from the uncorrected EDA signal. Therefore, this model not only exhibits a significant difference from the uncorrected EDA signal but also achieves artifact correction similar to the manually corrected signal. Thus, the model effectively accomplishes artifact removal in the EDA signal and makes novel contributions in this field.

Finally, these tools were applied in ER and depression recognition case studies. Both data sources were processed and several features were extracted to perform a classification task. The results obtained showed a fundamental need of the processing of these signals to find information and patterns by the ML models in the classification task. The use of biosignals represents different sources that could be profitable to optimize any kind of AfC experimentation. Several different but complementary type of information could be extracted from each signal. Moreover, this work also shows the necessity and the advantages in automatizing the signal processing of these kind of signals, reducing the time spent in the experimentation. Therefore, the optimization in the study of signal processing is a task necessary in the AfC field.

6.3 Virtual humans for emotion elicitation and mental health assessment

This work demonstrates how the construction and use of a VH is something that, despite its complexity, is currently feasible. Additionally, this work has shown how real-time conversations between humans and VH is possible, extracting valuable information from various data sources. The results exposed in this thesis show how the VH could achieve a high degree of naturalness and realism during the conversation. Moreover, it is demonstrated that VHs can serve as tools to aid ER and, especially, the diagnosis of mental illnesses. All in all, the results presented in this thesis represents a breakthrough in the research involved with the VHs.

The experimentation performed in this work, developed different VHs with various emotional states. It is demonstrated that in some cases, the use of multiple emotional states of VHs achieves a more comprehensive prediction, while in other instances, the use of a single VH can yield a better prediction. This opens up possibilities for contextualizing VHs, as different contexts can be specified based on the same emotional state to optimize ER. Furthermore, this thesis also demonstrates that robust identification of depressive patients can be achieved through human-VH interaction. This particularly opens the door to use this technology as a support tool for medical diagnosis, through an interaction similar to that of a specialist with a patient. Future experiments can explore this methodology and develop it for other types of mental illnesses.

This thesis represents an initial step of a path that can only improve. The development of technologies used for VH construction is continuously advancing. This will, therefore, allow for a much smoother interaction between human and VH. Greater realism in the VH, not only in appearance but also in voice, thus enabling better emotion elicitation will be achieved. Finally, LLM are currently experiencing rapid improvements. Increasingly, these models can generate a greater number of words and handle much larger contexts. This advancement will enable the specification of the VH emotional state in more detail, as well as achieving much more realistic and contextually coherent responses. This could mark the beginning of research whose main focus is the contextualization of VHs to obtain the highest score possible in ER. Despite being more complex, such research could bring about changes to the experiment itself through only a simple text input, yielding very different results and findings.

6.4 Future work

In this work, various avenues are opened, such as the exploration of algorithms related to ET and EDA, the implementation and improvement of different modules for the VH or the use of it in other type of experimentations. Different aspects can be further explored based on the findings presented in this thesis.

In the case of ET, the use of different algorithms and their adaptation to VR, if necessary, is a very interesting continuation of the study outlined in this thesis. Other calibration methodologies for ET algorithms can be explored using alternative algorithms, along with comparing their performance. Algorithms employing unsupervised ML can also be investigated to gain a better understanding of the subject's AoI and fixations performed during the experimentation. On the other hand, proposing different calibrations in a comprehensive study, where the subsequent validation of the calibration is examined, is particularly intriguing. This approach can yield valuable insights into the algorithm and parameters to be used.

Related to the study of EDA signal conducted in this work, the design and implementation of new architectures in this field is a first point to explore. Even the use of transformer layers to enhance the detection model or create a new regression model is something achievable today. The overall improvement of the model is a task that will require various novel resources from ML and DL, which are becoming more easily applicable every day. The investigation into model enhancement can also become a very extensive task that explores various research domains in the signal processing and AI fields.

Regarding the development of the VH, there are numerous extensions to consider. Firstly, updating the VH with various AI models that are emerging daily is something that should be explored. Additionally, more modules can be added to the VH, such as a real-time sentiment analysis. This module would identify the emotions in the subject's message, influencing the VH response, and based on it, the tone of the message could change. The introduction of a camera that can take photos or real-time videos of the subject, to analyze their movement and study the level of interest, is another aspect to explore. In conclusion, concerning the enhancement of the VH, there are numerous modules that can be improved and added to enhance interaction with humans increasing drastically the amount of information that could be extracted.

Lastly, all the points outlined for further investigation could potentially enhance the ER or the recognition of depressive patients studied in this work. Improvements in signal processing algorithms or the VH are aspects that can directly enhance the results of this thesis. Moreover, other signals widely used in ER or the detection of depressive subjects, such as ECG or EEG can be further studied. Throughout this work, various data sources have been explored to gather as much information as possible for performing classification tasks. In the case of these two signals, they are extensively used and investigated in different experiments. This exploration can enhance the detection of emotions or mental illnesses in conjunction with the use of VHs. The study of these signals, in this case, requires the development of automatic processing methods that allow their future real-time application. This opens the door to the development of calibration algorithms or artifact removal algorithms that may lead to very interesting results.

6.5 Future applications of the framework

The future applications of the proposed framework in this thesis, span a diverse array of domains. It has several applications in different fields but the main ones could be focused in the development of medical assessment for mental illness. For example, one potential application is the integration of this system into primary care settings to facilitate the early detection and stratification of depression, enabling the delivery of personalized medicine. This type of system would also avoid collapse in primary care, especially psychiatrists and psychologists. Moreover, the framework could be deployed as a home monitoring tool to detect early signs of depression relapse, allowing for timely intervention before the illness escalate. Notably, the system's automated functionality makes it well-suited for home use, allowing the removal of certain modules according with the services of the patient. Additionally, the framework holds potential for addressing a spectrum of mental health disorders beyond depression, including post-traumatic stress disorder, bipolar disorder, and borderline personality disorder. Furthermore, it could be adapted for use in supporting neuro-divergent individuals, aiding in the assessment and enhancement of social skills among those on the autism spectrum.

Furthermore, the framework holds potential for application in other contexts such as education or professional field. It could function as a valuable resource for training and refining social and communication skills in both, academic and professional environments. As an educational tool, it could augment traditional teaching methods by integrating into lesson contexts provided by instructors, even outside the school. In professional settings, the framework could be developed for team-building initiatives, conflict resolution strategies, and leadership programs. Insights gleaned from analyzing emotional dynamics and social interactions could offer invaluable perspectives on group organization, fostering more efficient and cohesive work environments.

Chapter 7

Conclusion

This thesis presents several novel contributions in the realm of signal analysis and affective computing (AfC) tasks. The primary aim of this research is to develop automated tools capable of eliciting and recognizing emotions within an AfC experimentation. To achieve this objective, various technologies have been developed and investigated. Initially, the focus was on extending an eye-tracking (ET) fixation identification algorithm from 2D to 3D in virtual reality (VR), incorporating head movement into the algorithm. Additionally, a novel methodology was introduced to analyze the optimal set of thresholds for the ET algorithm in VR. Another significant aspect of this study involves the analysis of electrodermal activity (EDA) signals. Specifically, an automatic artifact correction method was developed, yielding results comparable to manual corrections performed by experts. Leveraging advancements in both ET and EDA signals, automated emotion recognition became feasible. Subsequently, an AfC experiment was designed to elicit emotions, involving the development of a virtual human (VH) based on cutting-edge artificial inteligence (AI) technologies. This VH is capable of eliciting emotions through real-time, open-ended voice conversations with human subjects. The VH ability to elicit emotions in humans during these conversations was evaluated, alongside the identification of emotions expressed by the VH itself. Finally, all these tools and insights were integrated into a comprehensive experimentation framework. This framework facilitates the automatic processing of ET and EDA signals for emotion recognition (ER) and depression assessment, leveraging advanced statistical methods such as machine learning (ML).

The presented insights offer a comprehensive understanding of psycho-physiological responses to AI-based intelligent system for social emotion elicitation, paving the way for improved practices across various domains. The findings have the main implications in health and psychology, where this tool could help in the recognition and treatment of other mental illnesses or in the early diagnosis of them, aiding therapies such as phobia recovery or training in the educational and professional field. However, developing ER models in VR for extrapolation to other environments requires additional investigation. This calls for a new sub-field in AfC, necessitating future studies with larger datasets and participant numbers in various immersive settings.

In conclusion, emotions play a critical role in our daily lives, so an understanding and recognition of emotional responses is crucial for human research. We believe that the framework proposed can revolutionize emotion elicitation and recognition experiments, and will impact transversely in many areas of research, opening new opportunities for the scientist community.

Research Activities

Journal papers

Llanes-Jurado, J., Gómez-Zaragozá, L., Minissi, M. E., Alcañiz, M., & Marín-Morales, J. (2024). Developing conversational Virtual Humans for social emotion elicitation based on large language models. Expert Systems with Applications, 246, 123261. https://doi.org/10.1016/j.eswa.2024.123261.

Carrasco-Ribelles, L. A., **Llanes-Jurado, J.**, Gallego-Moll, C., Cabrera-Bean, M., Monteagudo-Zaragoza, M., Violán, C., & Zabaleta-del-Olmo, E. (2023). Prediction models using artificial intelligence and longitudinal data from electronic health records: A systematic methodological review. Journal of the American Medical Informatics Association, 30(12), 2072–2082. https://doi.org/10.1093/jamia/ocad168

Llanes-Jurado, J., Carrasco-Ribelles, L. A., Alcañiz, M., Soria-Olivas, E., & Marín-Morales, J. (2023). Automatic artifact recognition and correction for electrodermal activity based on LSTM-CNN models. Expert Systems with Applications, 230, 120581. https://doi.org/10.1016/j.eswa.2023.120581.

de-Juan-Ripoll, C., **Llanes-Jurado J.**, Chicchi Giglioli, I. A., Marín-Morales, J. & Alcañiz, M. (2021). An Immersive Virtual Reality Game for Predicting Risk Taking through the Use of Implicit Measures. Applied Sciences, 11(2), 825. https://doi.org/10.3390/app11020825

de-Juan-Ripoll, C., Chicchi Giglioli, I. A., **Llanes-Jurado, J.**, Marín-Morales, J., & Alcañiz, M. (2021). Why Do We Take Risks? Perception of the Situation and Risk Proneness Predict Domain-Specific Risk Taking. Frontiers in psychology, 12, 562381. https://doi.org/10.3389/fpsyg.2021.562381

Llanes-Jurado, J., Marín-Morales, J., Guixeres, J., & Alcañiz, M. (2020). Development and calibration of an eye-tracking fixation identification algorithm for immersive virtual reality. Sensors, 20(17), 4956.
https://doi.org/10.3390/s20174956.

Conference publications

Gómez-Zaragozá, L., Minissi, M.E., **Llanes-Jurado, J.**, Altozano, A., Alcañiz Raya, M., Marín-Morales, J. (2023). Linguistic Indicators of Depressive Symptoms in Conversations with Virtual Humans. In: Camarinha-Matos, L.M., Boucher, X., Ortiz, A. (eds) Collaborative Networks in Digitalization and Society 5.0. PRO-VE 2023. IFIP Advances in Information and Communication Technology, vol 688. Springer, Cham.
https://doi.org/10.1007/978-3-031-42622-3_37

Marín-Morales, J., **Llanes-Jurado, J.**, Minissi, M. E., Gómez-Zaragozá, L., Altozano, A., & Alcañiz, M. (2023). Gaze and head movement patterns of depressive symptoms during conversations with emotional virtual humans. In 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 1-8). Cambridge, MA, USA.
https://doi.org/10.1109/ACII59096.2023.10388134

Llanes-Jurado, J., Marín-Morales, J., Moghaddasi, M., Khatri, J., Guixeres, J. & Alcañiz, M. (2021). Comparing Eye Tracking and Head Tracking During a Visual Attention Task in Immersive Virtual Reality. Kurosu, M. (eds) Human-Computer Interaction. Interaction Techniques and Novel Applications. HCII 2021. Lecture Notes in Computer Science, 12763. Springer, Cham.
https://doi.org/10.1007/978-3-030-78465-2_3

Khatri, J., Moghaddasi, M., **Llanes-Jurado, J.**, Spinella, L., Marín-Morales, J., Guixeres, J., & Alcañiz, M. (2020). Segmentation of Areas of Interest Inside a Virtual Reality Store. In: Stephanidis, C., Antona, M. (eds) HCI International 2020 - Posters. HCII 2020. Communications in Computer and Information Science, vol 1225. Springer, Cham. https://doi.org/10.1007/978-3-030-50729-9_13

Khatri, J., Moghaddasi, M., **Llanes-Jurado, J.**, Spinella, L., Marín-Morales, J., Guixeres, J., & Alcañiz, M. (2020). Optimizing Virtual Reality Eye Tracking Fixation Algorithm Thresholds Based on Shopper Behavior and Age. In C. Stephanidis & M. Antona

(Eds.), HCI International 2020 - Posters (Vol. 1225, pp. 85-91). Springer, Cham. https://doi.org/10.1007/978-3-030-50729-9_9

Appendix A

Supplementary materials

This section provides supplementary materials detailing the results of the pretest performed to validate the faces designed for the emotional virtual humans (VH).

Pretest 1

A group of 56 subjects was recruited to participate in the experiment. The mean age of the subjects was 32.94 years, with a standard deviation (SD) of 11.45 years; the group included 30 males and 26 females.

To design four emotions (happy, relaxed, angry, and sad) and a neutral expression for two genders (male and female), 10 faces were created. An online survey was created to evaluate each face using the Self-Assessment Manikin in terms of arousal and valence, employing a Likert scale from 1 to 9.

Table A.1: Scores of arousal and valence for VH in pre-test 1

	Male				Female			
	Valence		Arousal		Valence		Arousal	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Neutral	4.89	0.82	3.84	1.73	5.11	0.82	3.64	1.99
Happy	6.18	1.21	3.73	1.85	6.18	1.54	4.61	1.91
Relaxed	4.71	1.04	3.36	1.81	5.11	1.25	3.82	1.88
Angry	2.93	1.50	6.27	1.65	2.84	1.75	7.00	1.33
Sad	2.46	1.64	5.91	2.09	2.48	0.95	6.04	1.79
Total	4.24	1.24	4.62	1.83	4.34	1.26	5.02	1.78

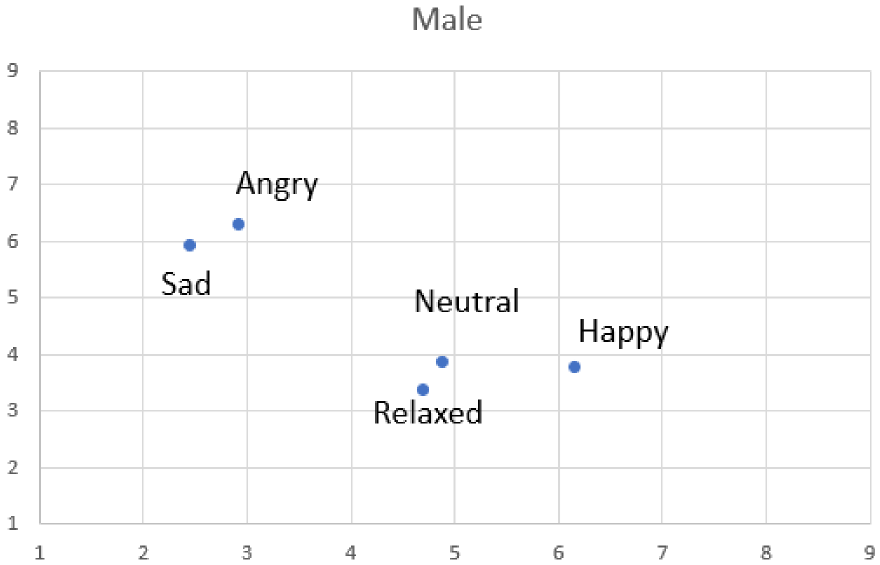


Figure A.1: Scores of the valence and arousal in male VH for subject self-assessment in pre-test 1

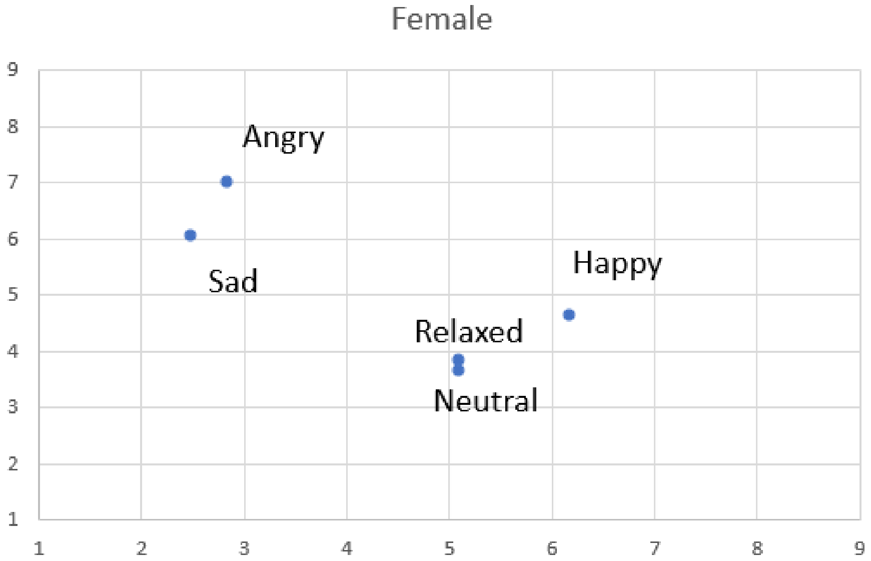


Figure A.2: Scores of the valence and arousal in female VH for subject self-assessment in pre-test 1

The main conclusions of the pretest for both genders were as follows:

- “Angry” was correctly placed in the quadrant denoting high arousal and negative valence.

- “Sad” was evaluated with higher arousal than expected. Corrections are needed to decrease arousal.

- “Relaxed” was very close to the “Neutral” face. Corrections are needed to increase positive valence.

- “Happy” was close to the correct quadrant, but corrections are needed to increase both arousal and positive valence.

- The “Neutral” face was close to the midpoint; therefore, no corrections are needed.

Pretest 2

A group of 46 subjects was recruited to participate in the experiment. The mean age of the subjects was 37.32 years, SD = 13.98, including 23 males and 23 females.

In this second phase, the 10 faces were modulated to ensure that each basic emotion was evaluated in the theoretical quadrant of the Circumplex Models of Affect, i.e., happy in high arousal and positive valence, relaxed in low arousal and positive valence, angry in high arousal and negative valence, and sad in low arousal and negative valence. An online survey was developed to evaluate each face using the Self-Assessment Manikin to measure arousal and valence, utilizing a Likert scale from 1 to 9.

Table A.2: Scores of arousal and valence for VH in pre-test 2

	Male				Female			
	Valence		Arousal		Valence		Arousal	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Neutral	4.91	0.89	3.91	1.66	5.28	0.86	3.57	1.67
Happy	6.83	1.22	3.61	2.07	6.61	1.56	4.63	1.77
Relaxed	5.67	0.87	3.35	1.72	6.20	1.36	3.80	1.98
Angry	2.59	1.50	7.04	1.26	2.63	1.34	6.33	1.61
Sad	2.70	1.24	5.37	1.74	2.74	1.39	5.04	1.84
Total	4.54	1.14	4.66	1.69	4.69	1.30	4.67	1.77

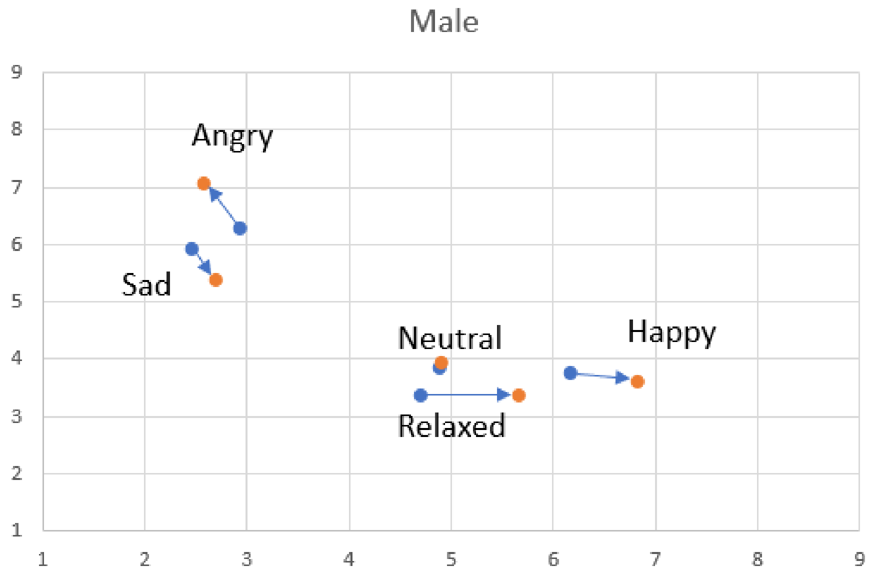


Figure A.3: Scores of the valence and arousal in male VH for subject self-assessment in pre-test 1 (blue dots) and pre-test 2 (orange dots)

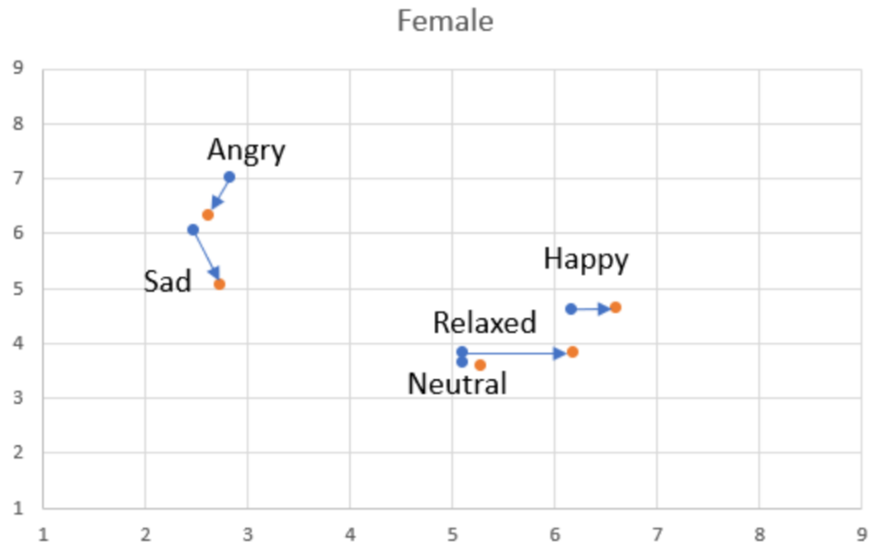


Figure A.4: Scores of the valence and arousal in female VH for subject self-assessment in pre-test 1 (blue dots) and pre-test 2 (orange dots)

The main conclusions of the pretest for both genders are as follows:

- The emotion of “anger” was successfully achieved.

- "Sad" faces showed a decrease in arousal but still experienced higher arousal than expected. However, the final values were close to 5.
- "Relaxed" faces notably increased in valence and are now in the correct quadrant.
- "Happy" faces increased in valence but did not see an increase in arousal. For females, arousal is close to 5, aligning with the theoretical quadrant; however, the male face is still distinctly placed in the quadrant of high arousal and low valence.

Final Conclusions

We successfully achieved neutral, angry, and relaxed faces. Sad faces experienced higher arousal, and happy faces showed lower arousal than expected. The results exhibit a bias to increase arousal in negative conditions and decrease it in positive ones. However, given that this is a static face that will be used dynamically during a conversation, where the face will be modulated by lip synchronization, we decided not to develop more extreme faces, as it can provoke uncanny valley reactions during conversations.

Final faces

The following screenshots are the final faces validated in the second pre-test and used in the experiment.

Neutral:

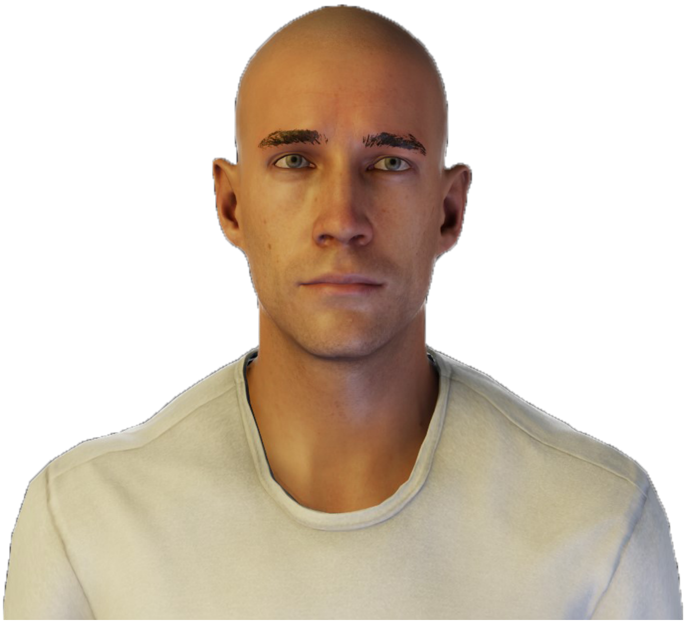


Figure A.5

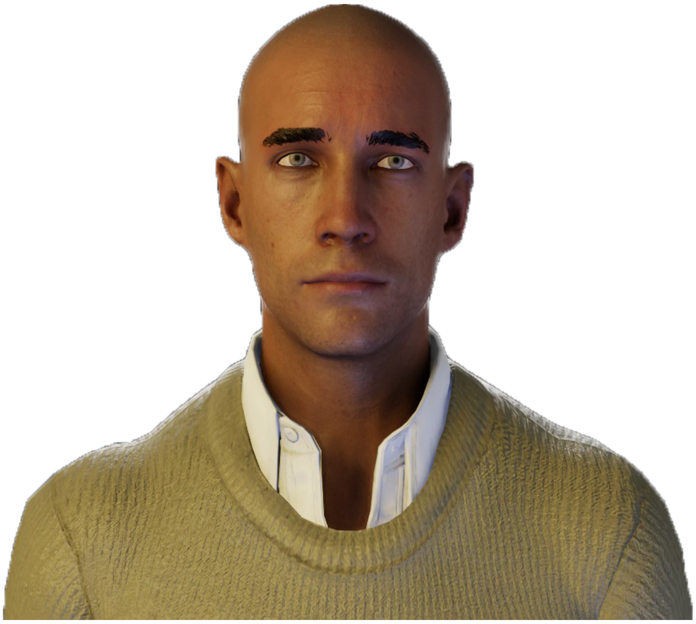


Figure A.6

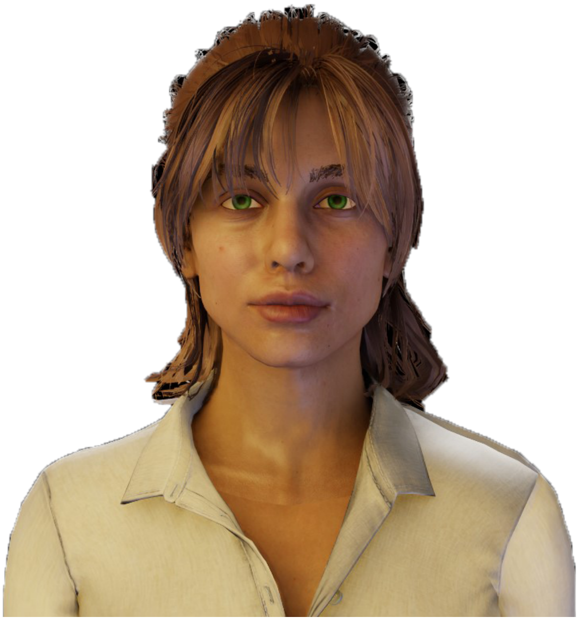


Figure A.7

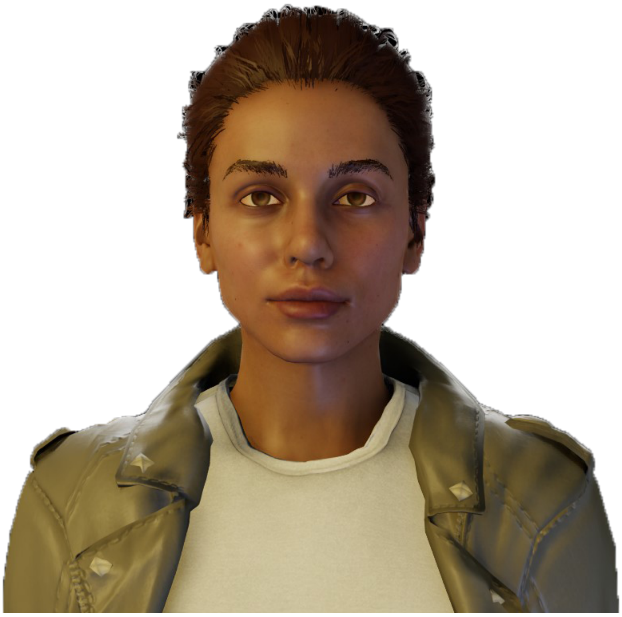


Figure A.8

Angry:

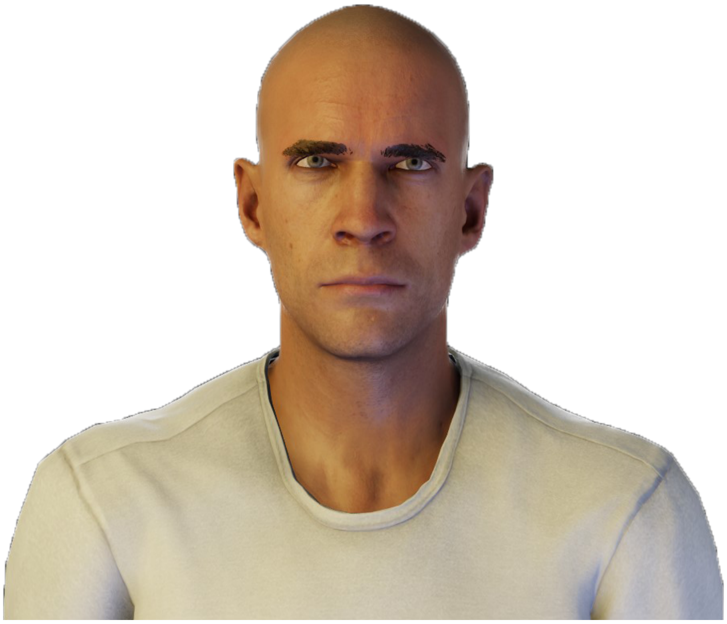


Figure A.9

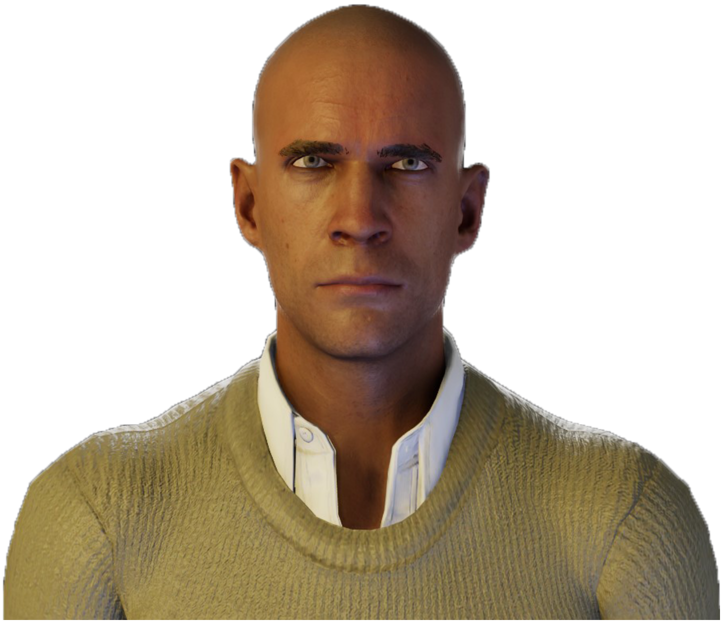


Figure A.10

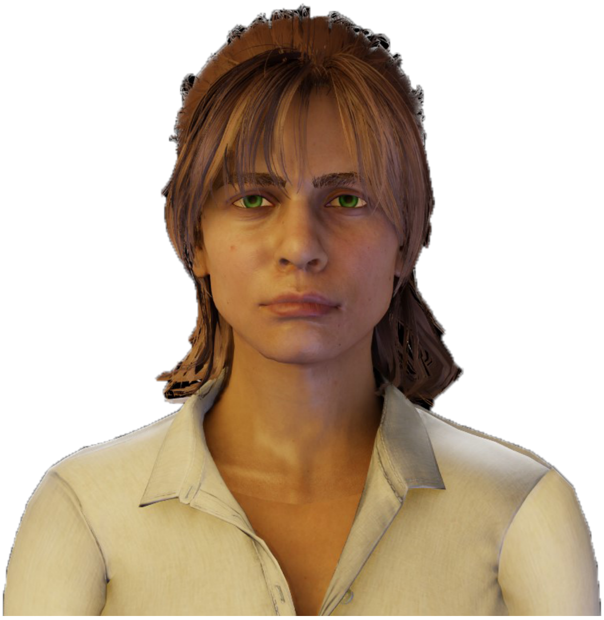


Figure A.11

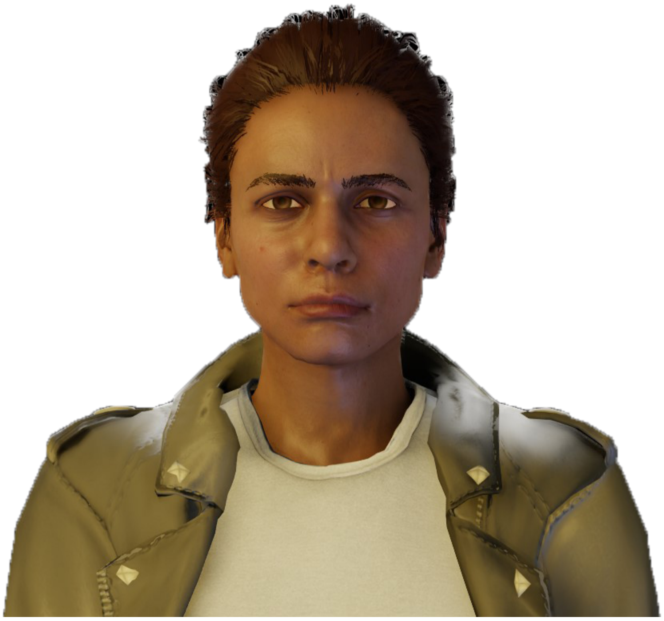


Figure A.12

Sad:

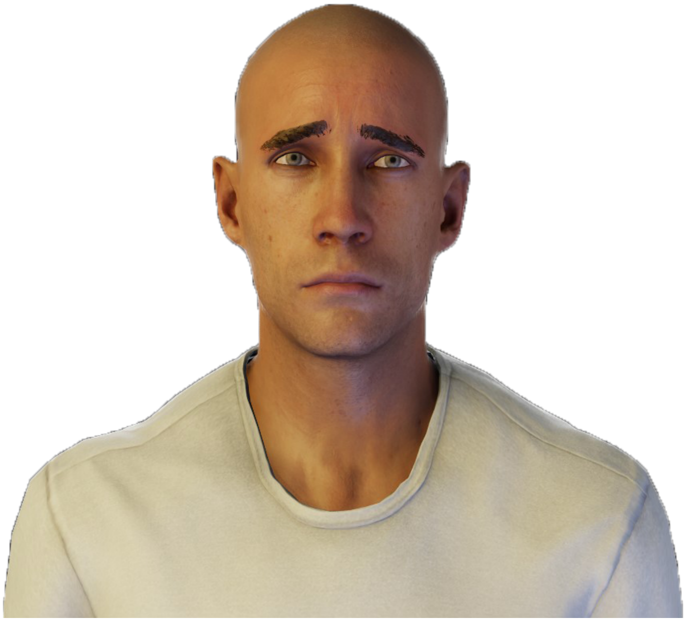


Figure A.13

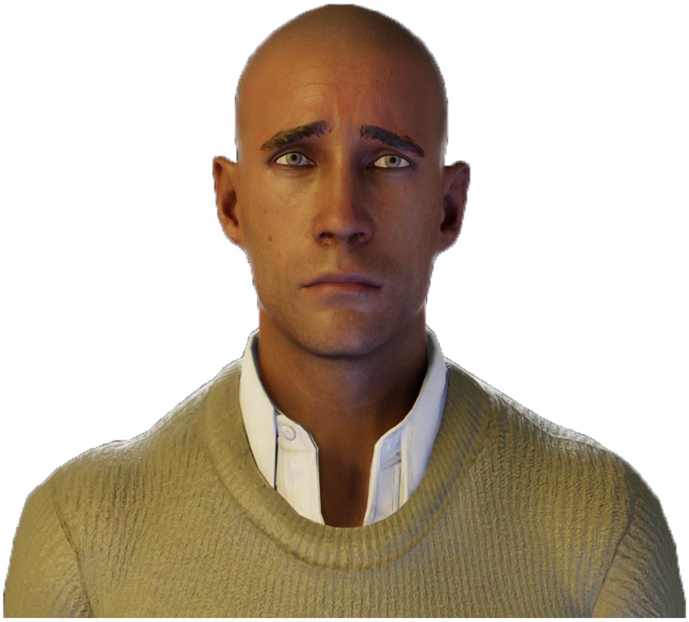


Figure A.14

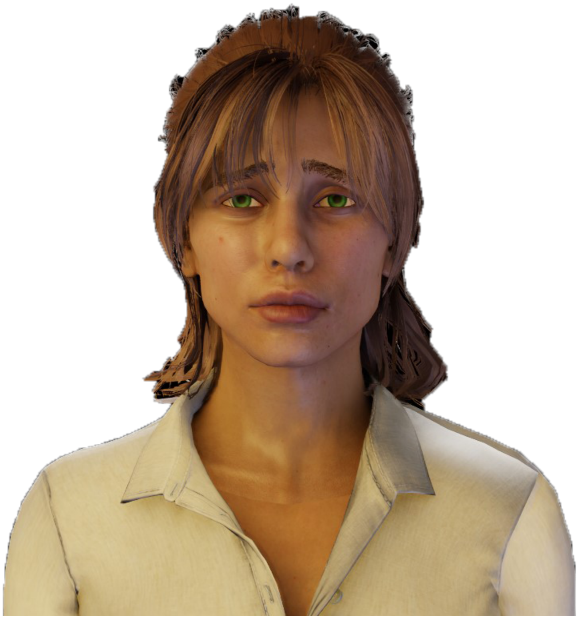


Figure A.15

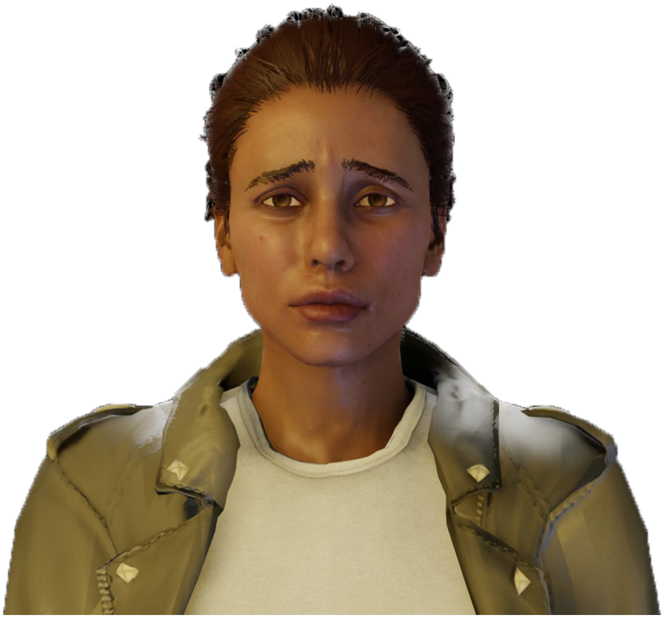


Figure A.16

Relaxed:

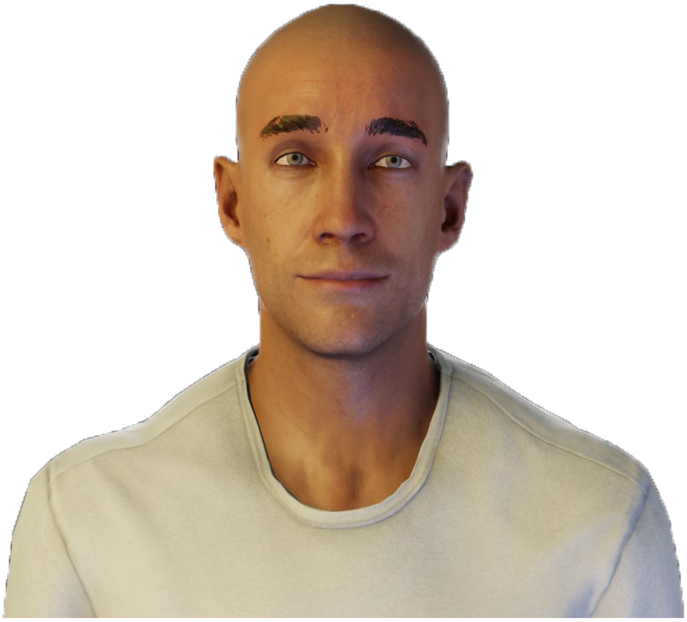


Figure A.17

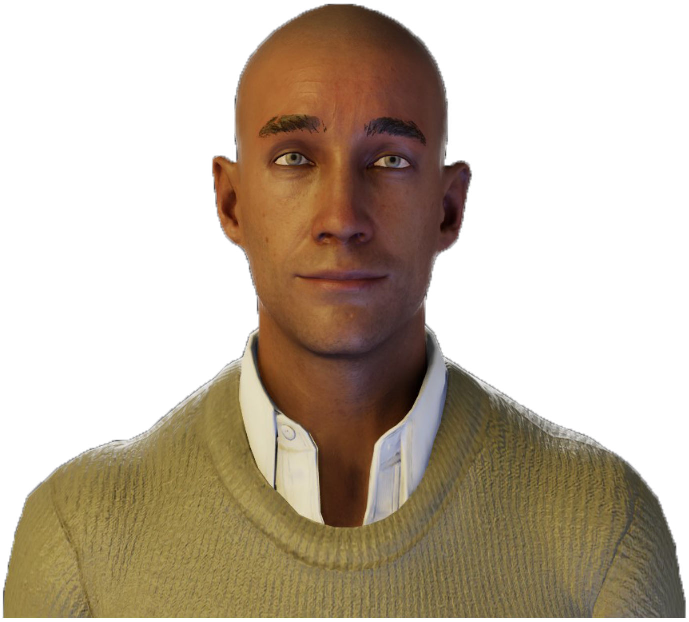


Figure A.18

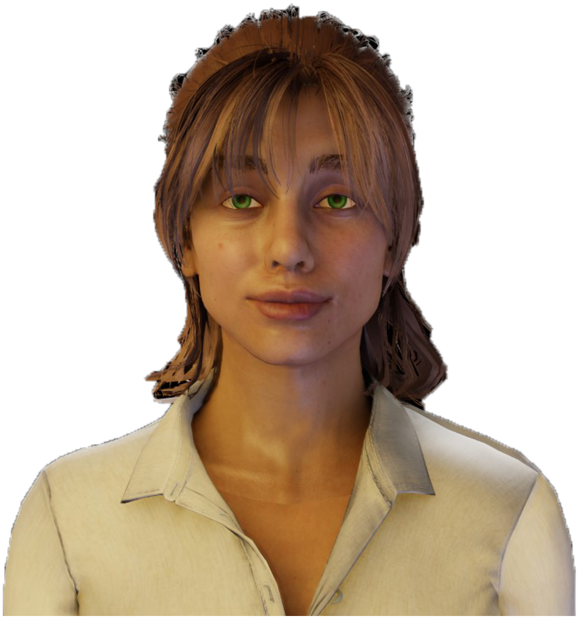


Figure A.19

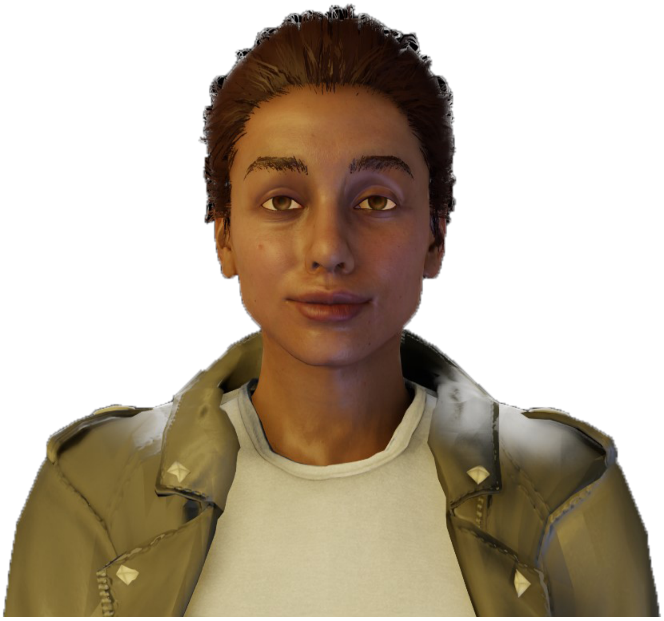


Figure A.20

Happy:

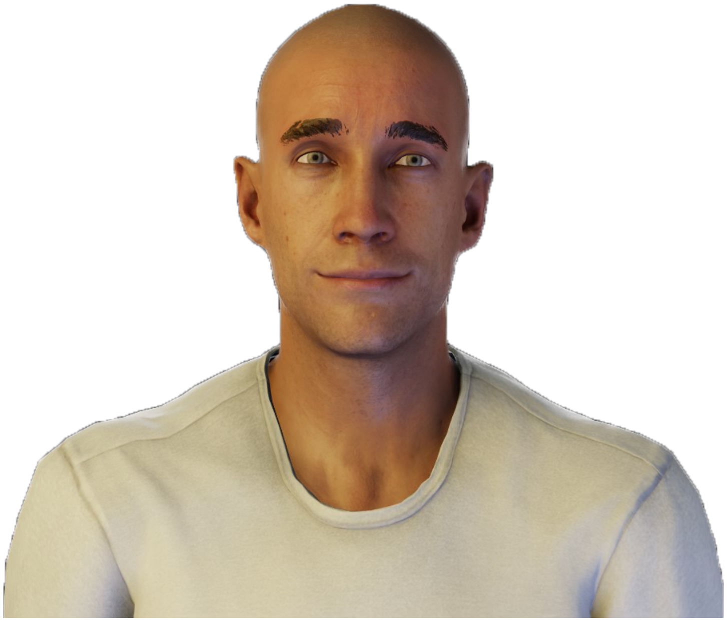


Figure A.21

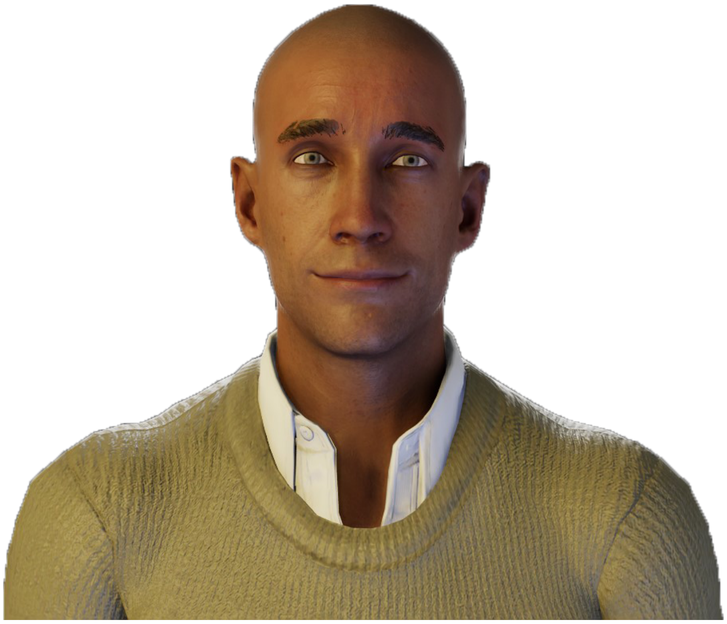


Figure A.22

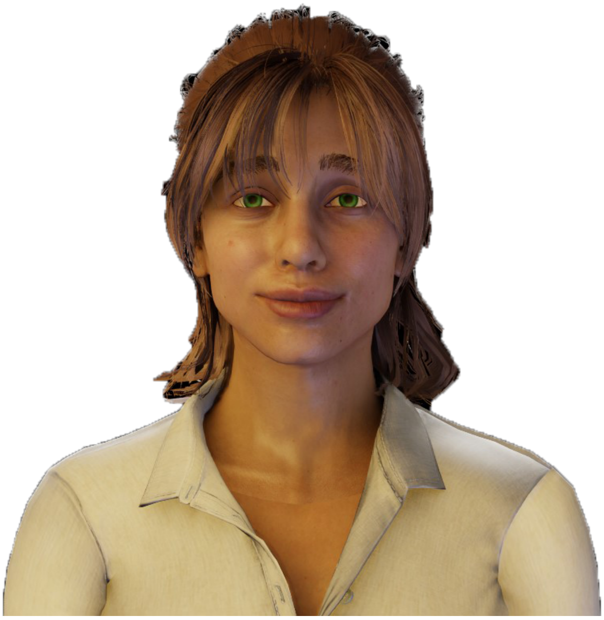


Figure A.23

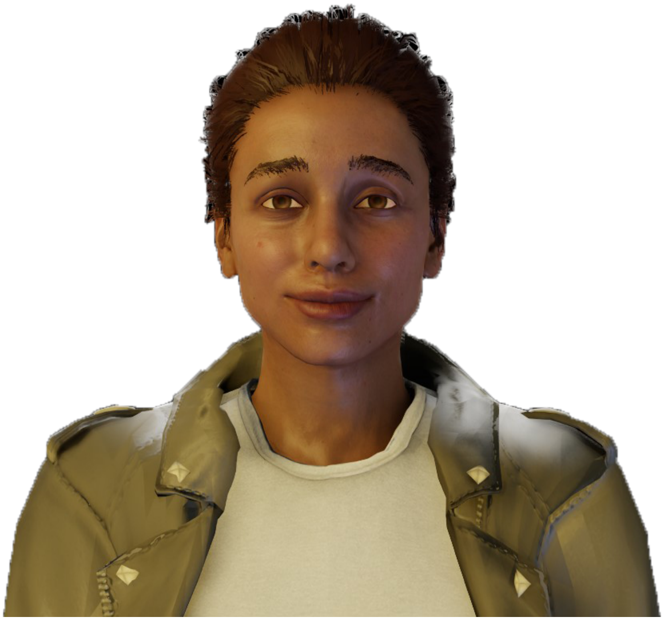


Figure A.24

Appendix B

Naturalness, Realism and Valence and Arousal in Human and VH graphical results

Naturalness

Figure B.1 shows the distribution over the test results for the prediction of naturalness for the different virtual human (VH) emotional states.

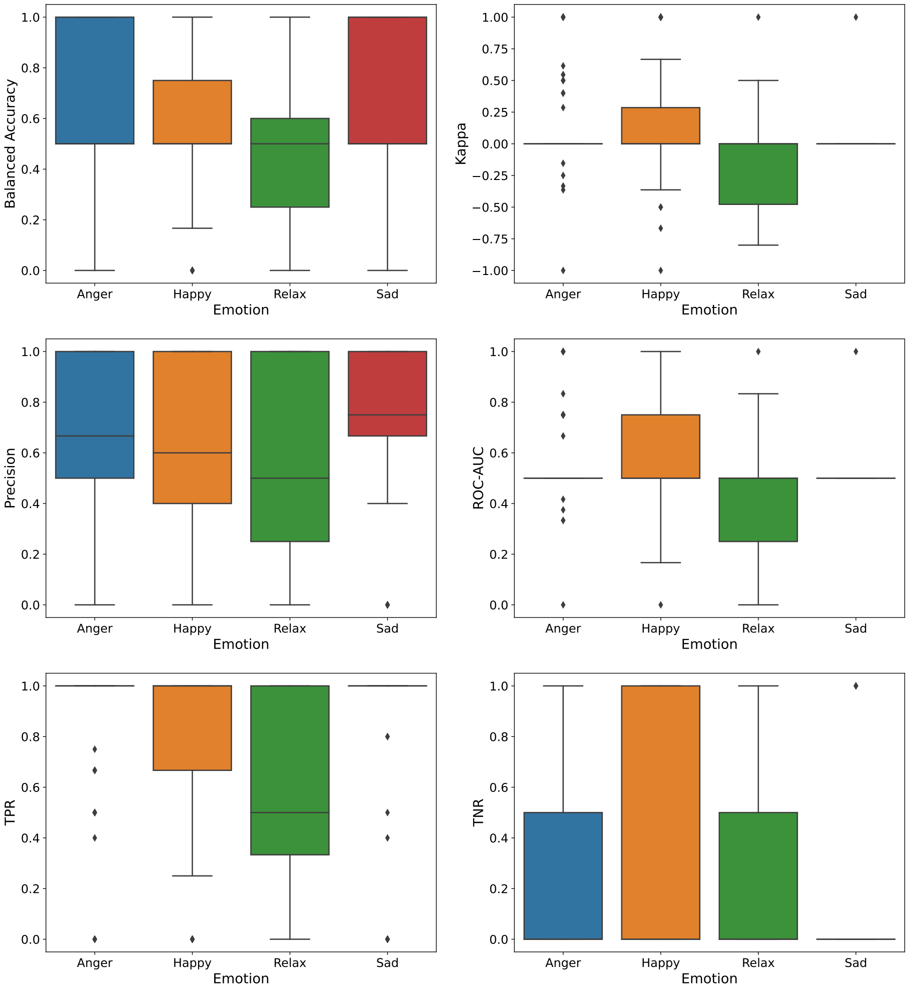


Figure B.1: Boxplots over the test distribution of the different scores measured for the Naturalness prediction. The metrcis are, from above to below and from right to left the following ones: Accuracy, Cohen-Kappa, Precision, RoC, TPR and TNR.

Figure B.2 shows the AUC curve for the prediction of naturalness in the test set. There are compared the different curves obtained for each different emotional VH. It is also added the AUC curve of the general model.

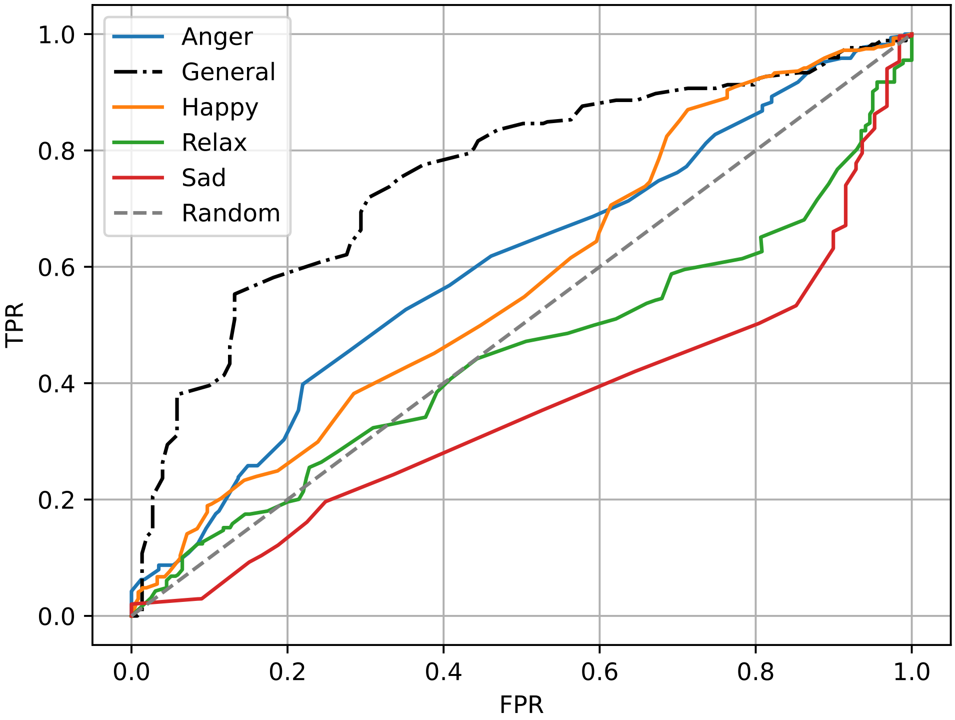


Figure B.2: AUC curve of the different models in terms of the VH emotional state for the Naturalness prediction. The general model is in black dashed line. The diagonal line in grey indicates the performance in the AUC curve of a random model.

Figure B.2 shows different type of curves. The curves corresponding to the emotional states sad and relax does not overpass the diagonal line which represents a random model. However, the states Happy and Anger are above the general model in most of the trajectory, being the Happy the one that is above mostly. The trajectory of the general model shows that are thresholds which are inconvenient for the model performance, but other set of thresholds could achieve a high performance of the model.

Figure B.3 shows the distribution of the naturalness target against five target randomly generated. Different scores are shown for the comparison of both distributions.

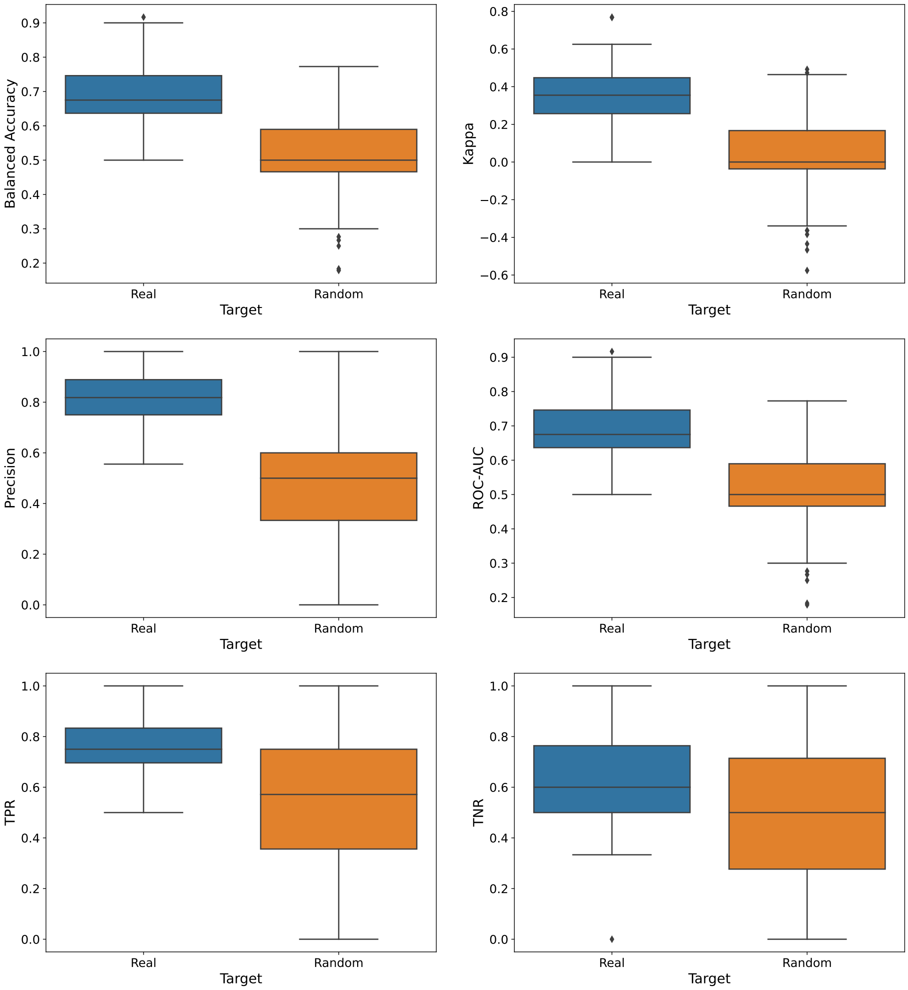


Figure B.3: Distributions of the different scores obtained for each prediction on the test set to check for overfitting in the naturalness models. In dark blue, we have the model prediction through the actual target, while in orange, we find the model prediction of the overfitting check.

Realism

Figure B.4 shows the distribution over the test results for the prediction of realism for the different VH emotional states.

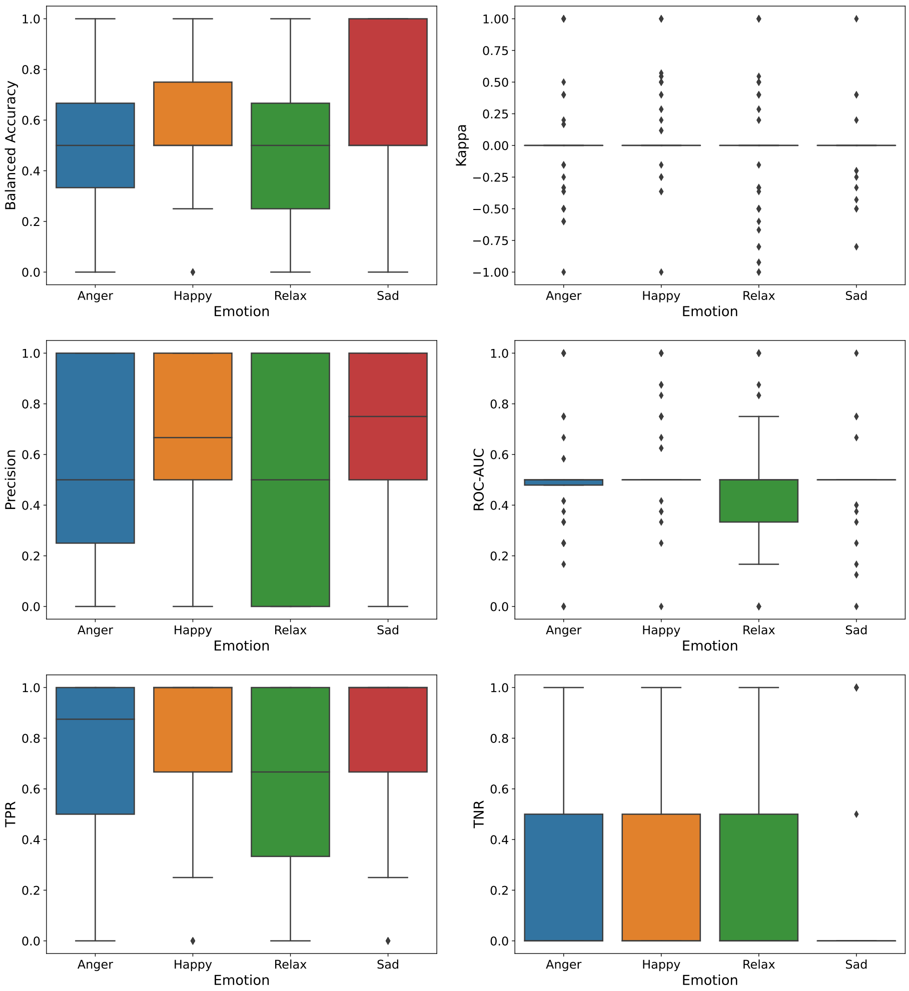


Figure B.4: Boxplots over the test distribution of the different scores measured for the realism prediction. The metrcis are, from above to below and from right to left the following ones: Accuracy, Cohen-Kappa, Precision, RoC, TPR and TNR.

Figure B.5 shows the AUC curve for the prediction of realism in the test set. There are compared the different curves obtained for each different emotional VH. It is also added the AUC curve of the general model.

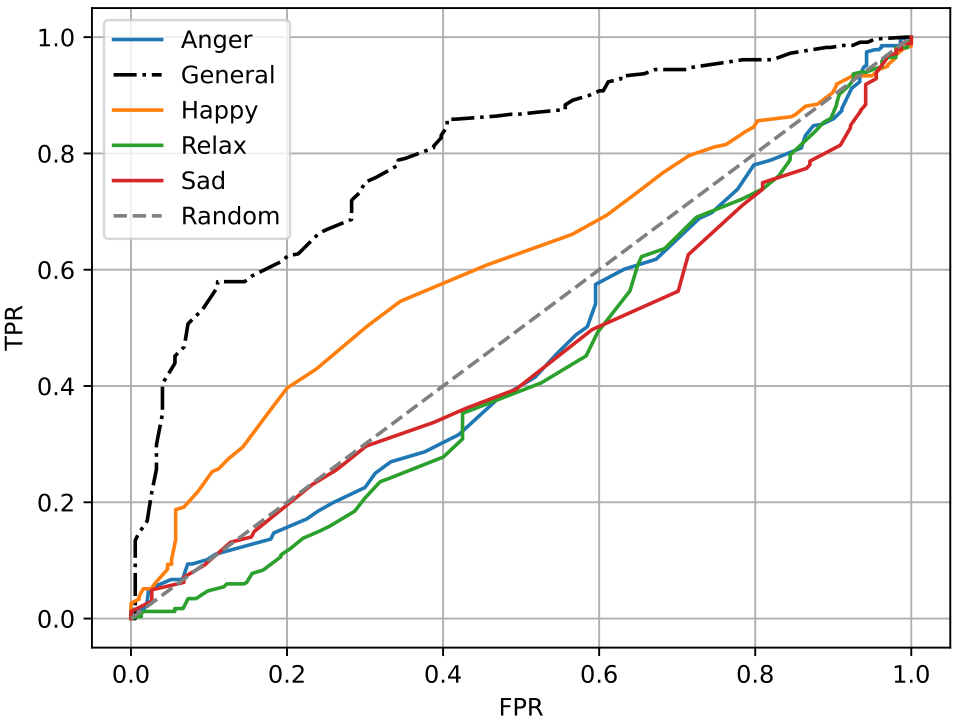


Figure B.5: AUC curve of the different models in terms of the VH emotional state for the realism prediction. The general model is in black dashed line. The diagonal line in grey indicates the performance in the AUC curve of a random model.

Figure B.5 shows different type of curves. The curves corresponding to the emotional states sad, relax and anger does not overpass the diagonal line which represents a random model in the AUC plot. However, the happy state is above the diagonal but below the general model in the whole trajectory.

Figure B.6 shows the distribution of the realism target against five target randomly generated. Different scores are shown for the comparison of both distributions.

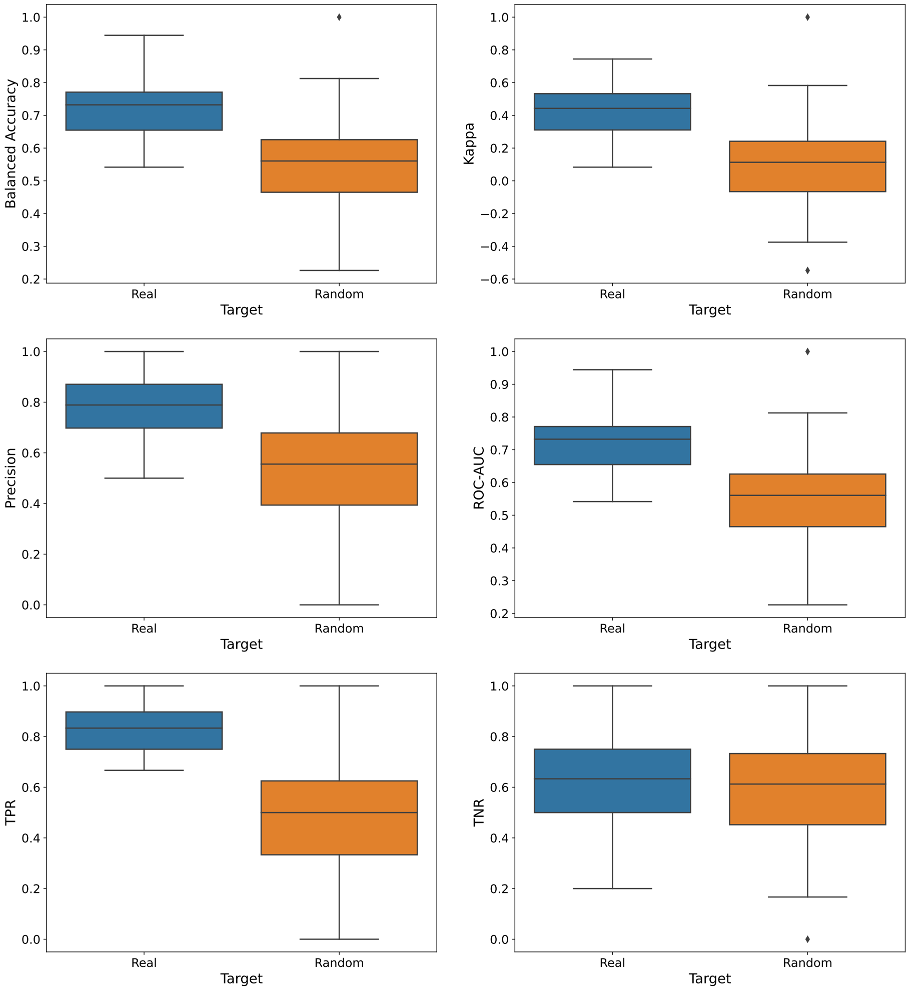


Figure B.6: Distributions of the different scores obtained for each prediction on the test set to check for overfitting in the realism model. In dark blue, we have the model prediction through the actual target, while in orange, we find the model prediction of the overfitting check.

Human Arousal

Figure B.7 shows the distribution over the test results for the prediction of human arousal for the different VH emotional states.

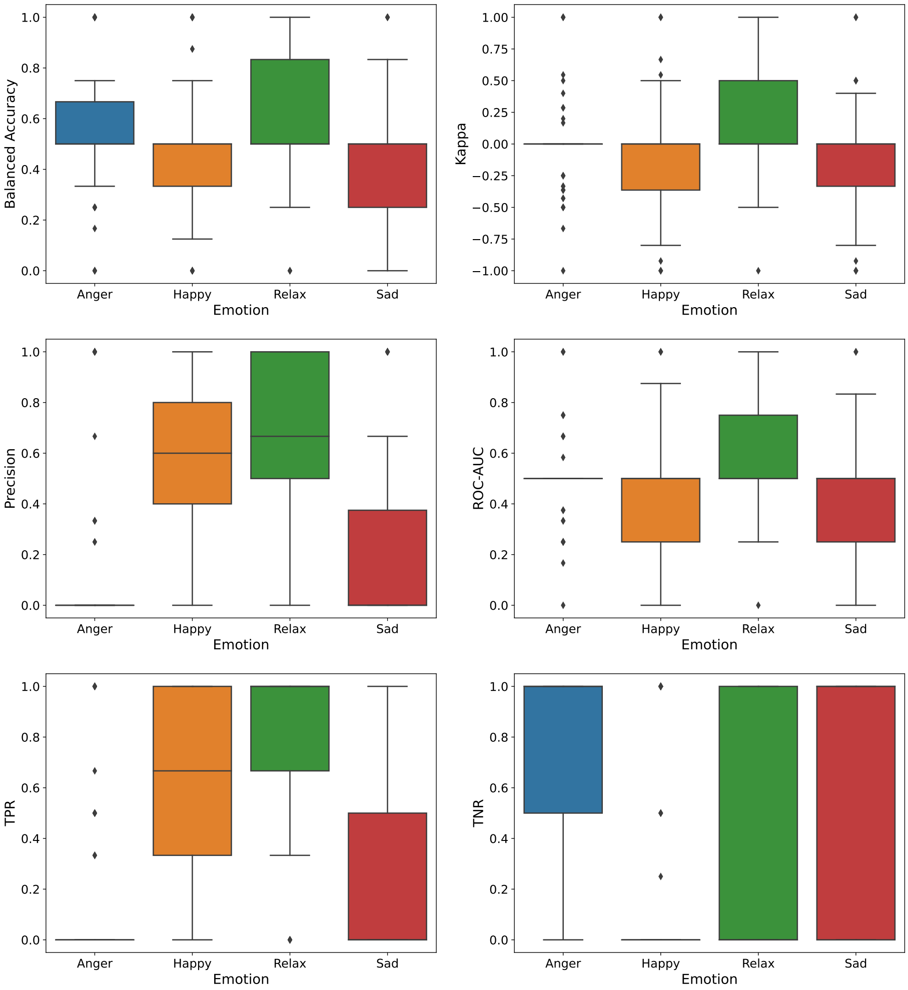


Figure B.7: Boxplots over the test distribution of the different scores measured for the human arousal prediction. The metrcis are accuracy, cohen-kappa, precision, ROC-AUC, TPR and TNR.

Figure B.8 shows the AUC curve for the prediction of human arousal in the test set. There are compared the different curves obtained for each different emotional VH. It is also added the AUC curve of the general model.

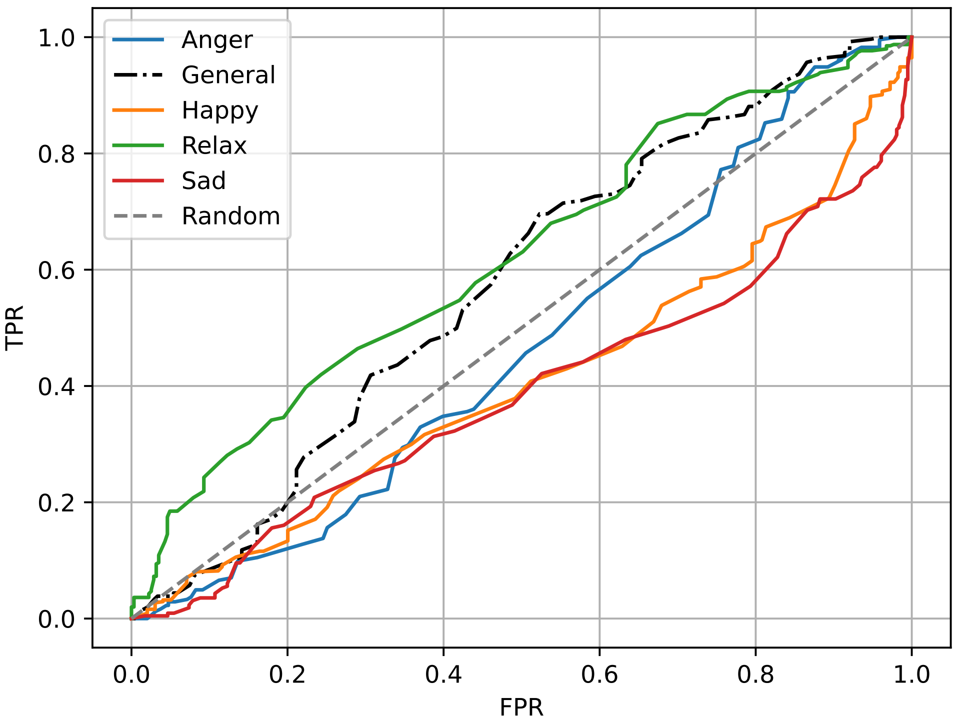


Figure B.8: AUC curve of the different models in terms of the VH emotional state for the human arousal prediction. The general model is in black dashed line. The diagonal line in grey indicates the performance in the AUC curve of a random model.

Figure B.5 shows different type of curves. The curves corresponding to the emotional states sad, relax and anger does not overpass the diagonal line which represents a random model in the AUC plot. However, the happy state is above the diagonal but below the general model in the whole trajectory.

Figure B.9 shows the distribution of the human arousal target against five target randomly generated. Different scores are shown for the comparison of both distributions.

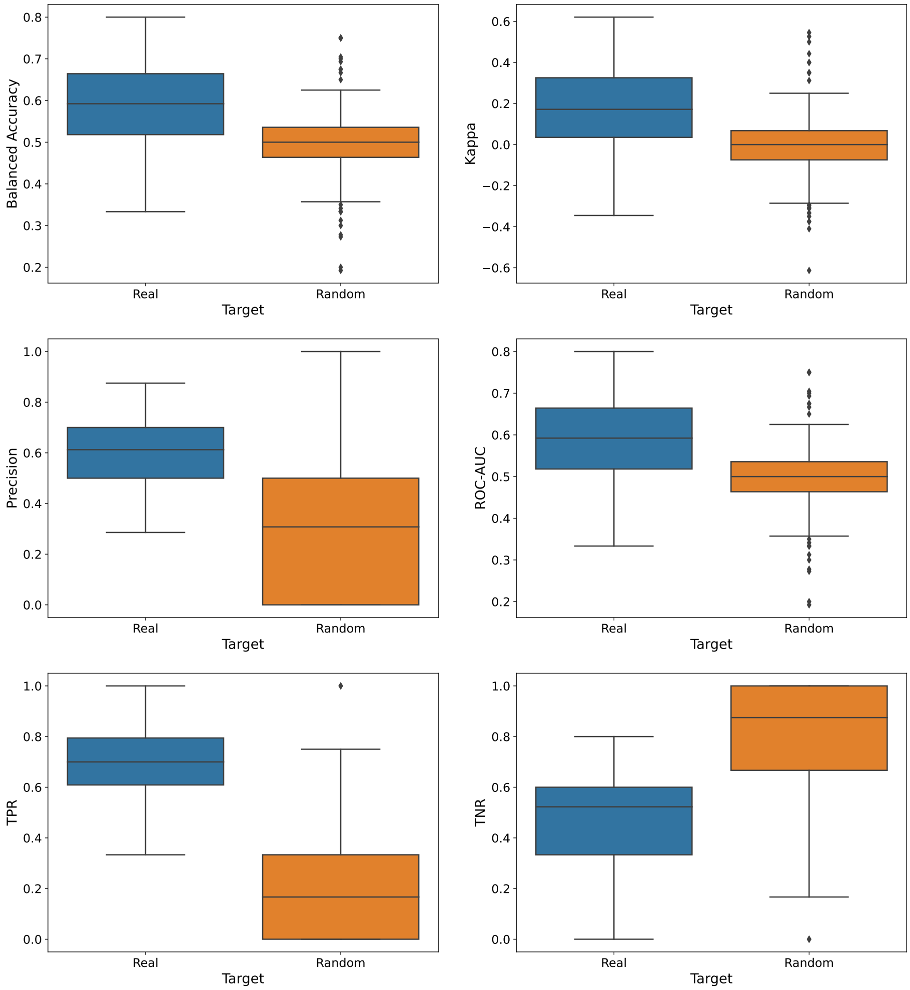


Figure B.9: Distributions of the different scores obtained for each prediction on the test set to check for overfitting in the human arousal model. In dark blue, we have the model prediction through the actual target, while in orange, we find the model prediction of the overfitting check.

Human Valence

Figure B.10 shows the distribution over the test results for the prediction of human valence for the different VH emotional states.

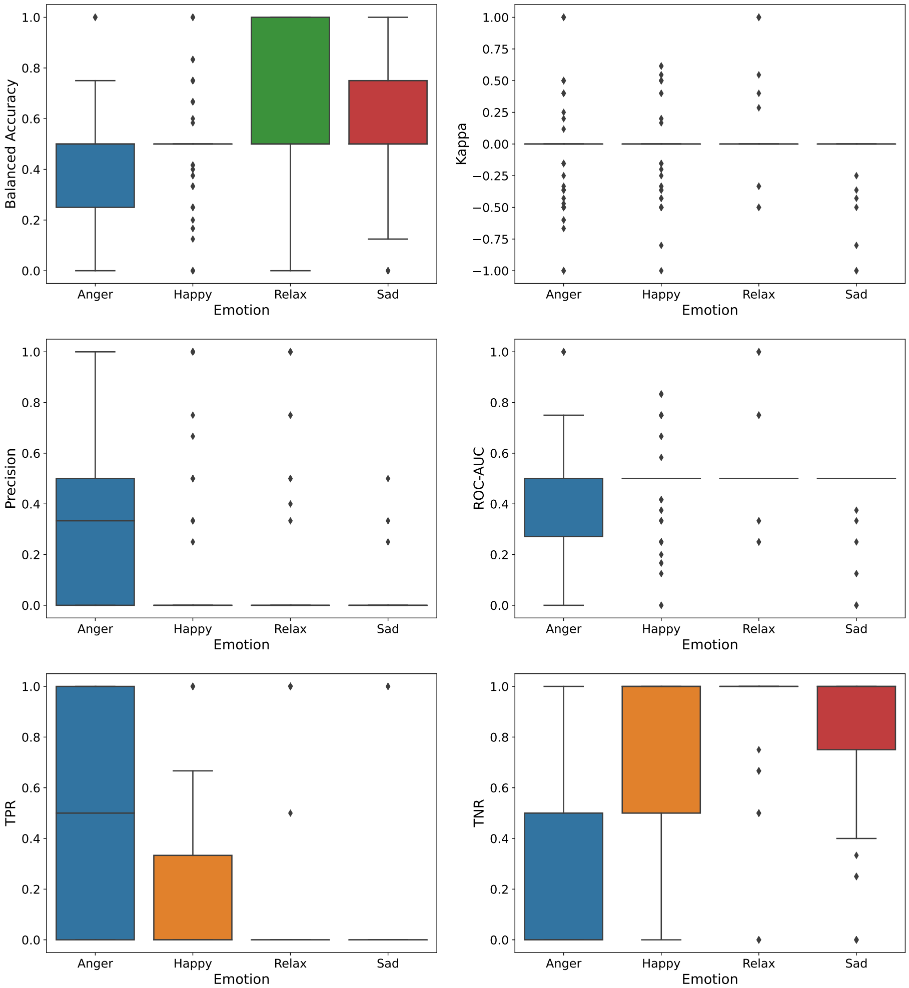


Figure B.10: Boxplots over the test distribution of the different scores measured for the human valence prediction. The metrcis are (a) Accuracy, (b) Cohen-Kappa, (c) Precision, (d) ROC-AUC, (e) TPR and (f) TNR.

Figure B.11 shows the AUC curve for the prediction of human valence in the test set. There are compared the different curves obtained for each different emotional VH. It is also added the AUC curve of the general model.

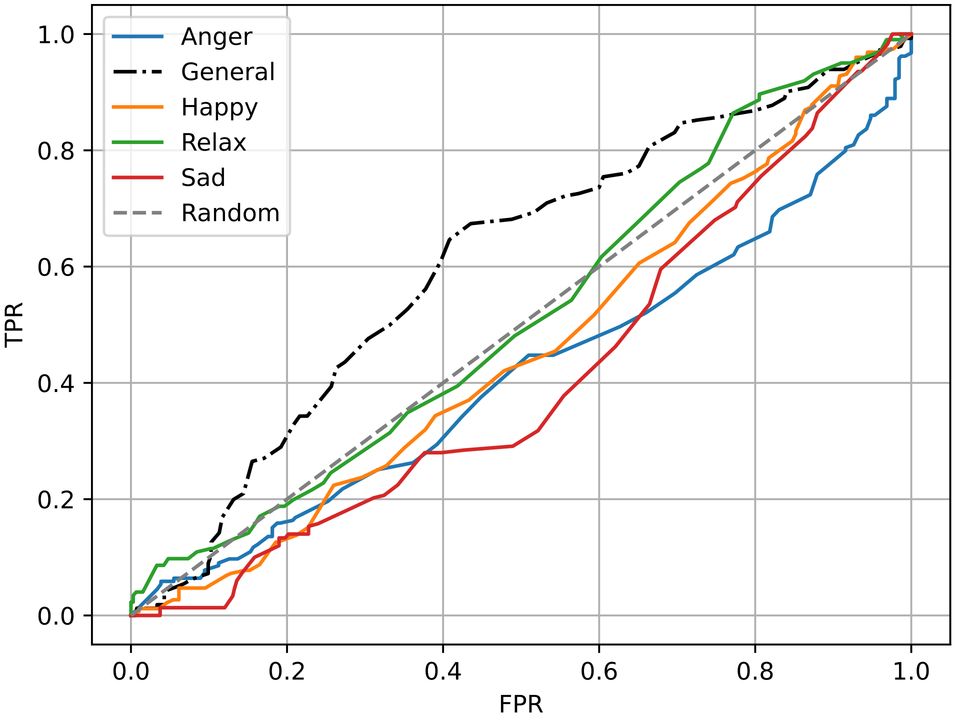


Figure B.11: AUC curve of the different models in terms of the VH emotional state for the human valence prediction. The general model is in black dashed line. The diagonal line in grey indicates the performance in the AUC curve of a random model.

Figure B.11 shows different type of curves. The curves corresponding to the emotional state sad, happy and anger does not overpass the diagonal line which represents a random model in the AUC plot. However, the relax could overpass the diagonal line in certain regions. All the emotional states are below the general model in the whole trajectory.

Figure B.12 shows the distribution of the human valence target against five target randomly generated. Different scores are shown for the comparison of both distributions.

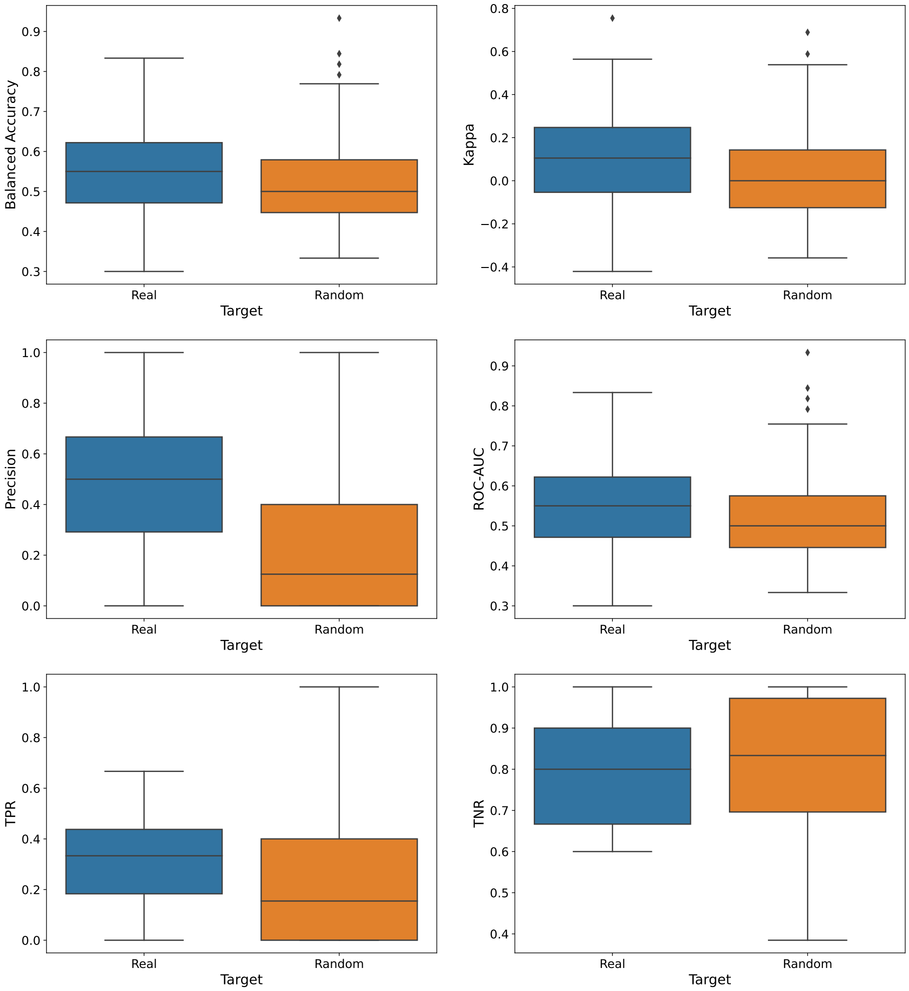


Figure B.12: Distributions of the different scores obtained for each prediction on the test set to check for overfitting in the human valence model. In dark blue, we have the model prediction through the actual target, while in orange, we find the model prediction of the overfitting check.

VH Arousal

Figure B.13 shows the distribution over the test results for the prediction of VH arousal for the different VH emotional states.

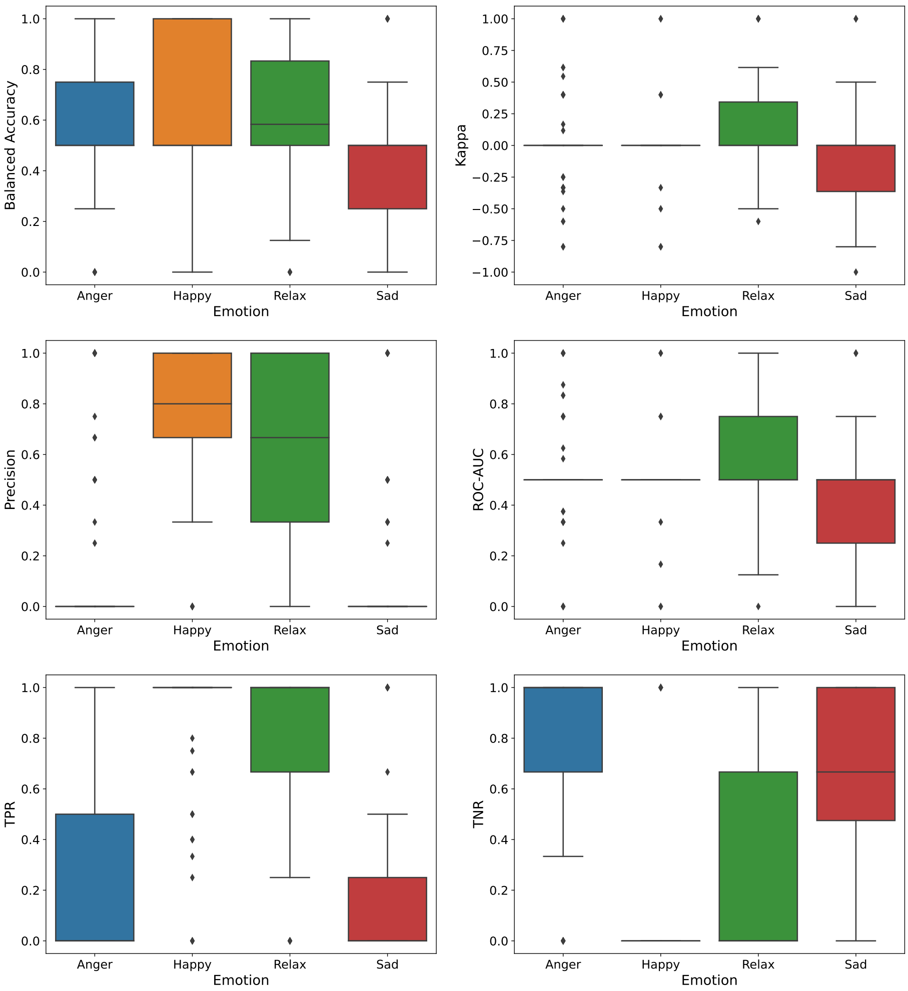


Figure B.13: Boxplots over the test distribution of the different scores measured for the VH arousal prediction. The metrcis are (a) Accuracy, (b) Cohen-Kappa, (c) Precision, (d) ROC-AUC, (e) TPR and (f) TNR.

Figure B.14 shows the AUC curve for the prediction of VH arousal in the test set. There are compared the different curves obtained for each different emotional VH. It is also added the AUC curve of the general model.

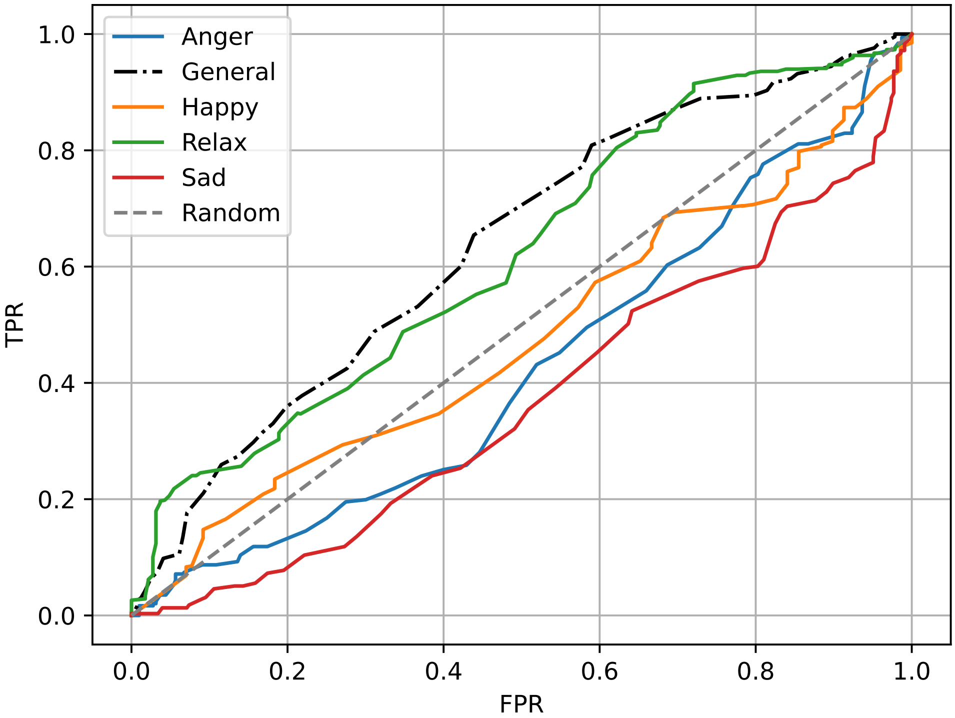


Figure B.14: AUC curve of the different models in terms of the VH emotional state for the VH arousal prediction. The general model is in black dashed line. The diagonal line in grey indicates the performance in the AUC curve of a random model.

Figure B.14 shows different type of curves. The curves corresponding to the emotional state sad, happy and anger does not overpass the diagonal line which represents a random model in the AUC plot. However, the relax state has a similar trajectory to the general model, which also is overpassed by the relax model in certain probability regions.

Figure B.15 shows the distribution of the VH arousal target against five target randomly generated. Different scores are shown for the comparison of both distributions.

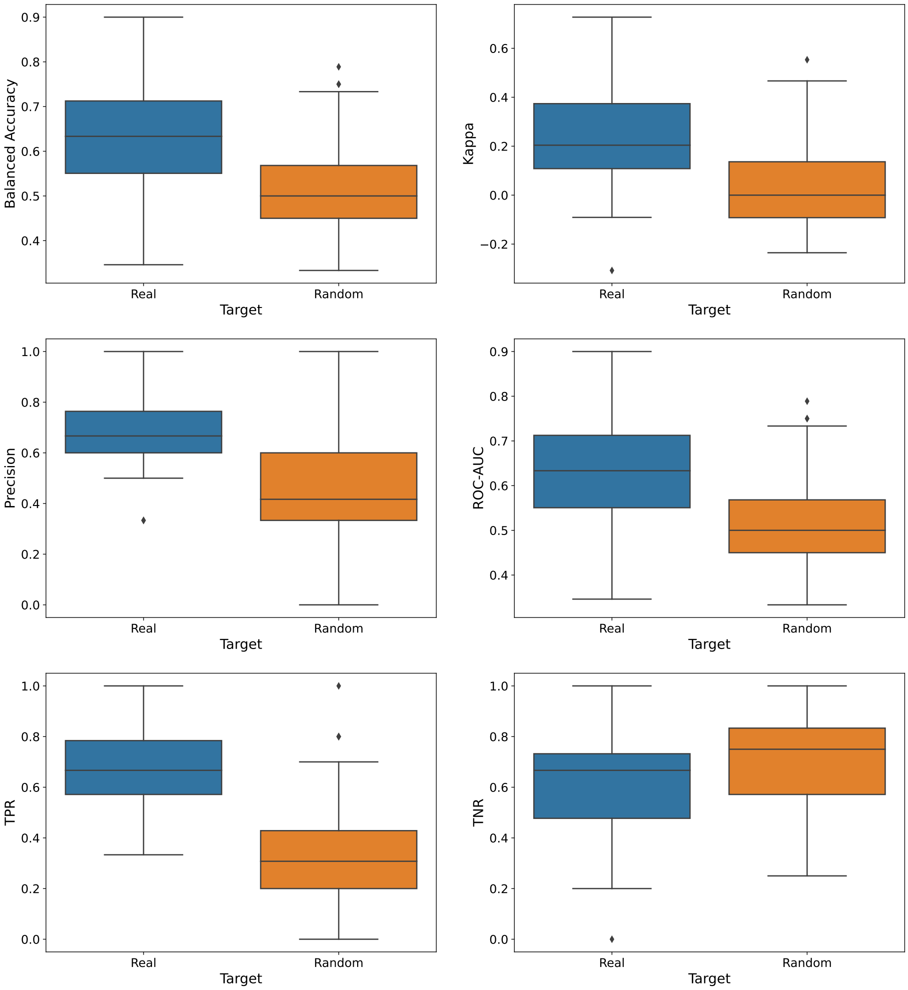


Figure B.15: Distributions of the different scores obtained for each prediction on the test set to check for overfitting in the VH arousal model. In dark blue, we have the model prediction through the actual target, while in orange, we find the model prediction of the overfitting check.

VH Valence

Figure B.16 shows the distribution over the test results for the prediction of VH valence for the different VH emotional states.

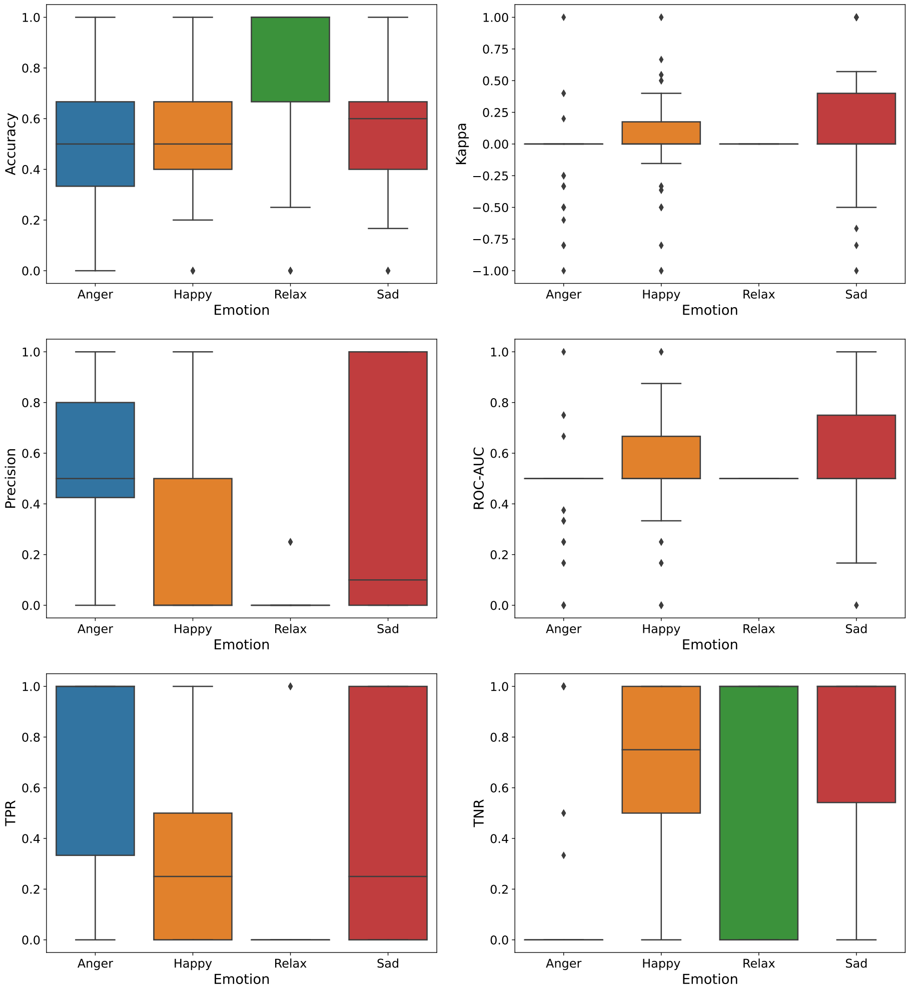


Figure B.16: Boxplots over the test distribution of the different scores measured for the VH valence prediction. The metrcis are (a) Accuracy, (b) Cohen-Kappa, (c) Precision, (d) ROC-AUC, (e) TPR and (f) TNR.

Figure B.17 shows the AUC curve for the prediction of VH valence in the test set. There are compared the different curves obtained for each different emotional VH. It is also added the AUC curve of the general model.

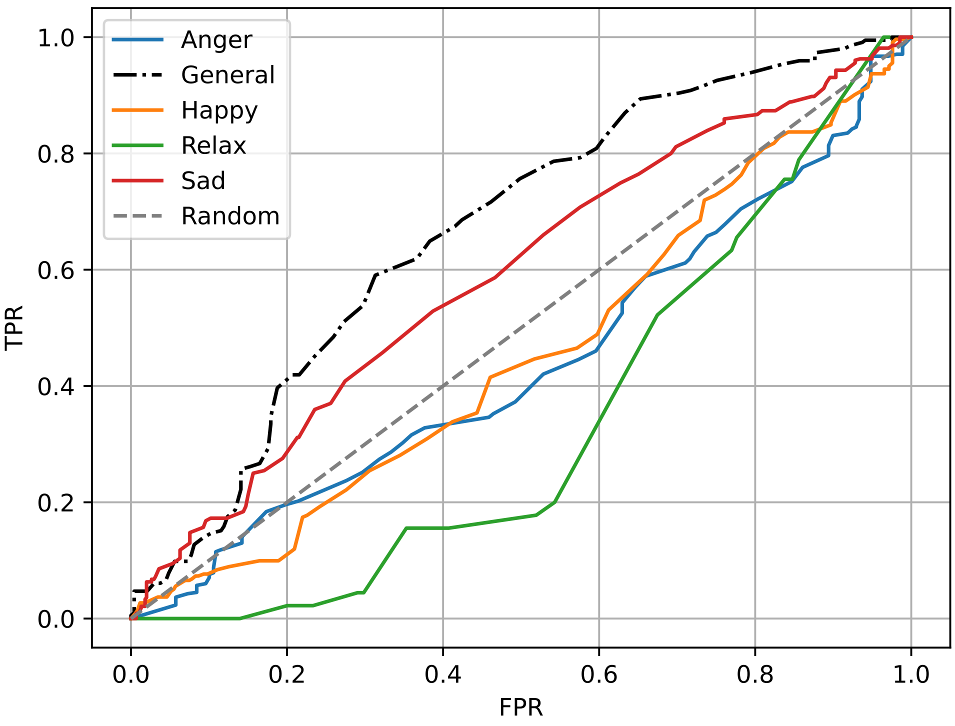


Figure B.17: AUC curve of the different models in terms of the VH emotional state for the VH valence. The general model is in black dashed line. The diagonal line in grey indicates the performance in the AUC curve of a random model.

Figure B.17 shows different type of curves. The curves corresponding to the emotional state relax, happy and anger does not overpass the diagonal line which represents a random model in the AUC plot. However, the sad state overpass this line in the whole trajectory but always below the general model.

Figure B.18 shows the distribution of the VH valence target against five target randomly generated. Different scores are shown for the comparison of both distributions.

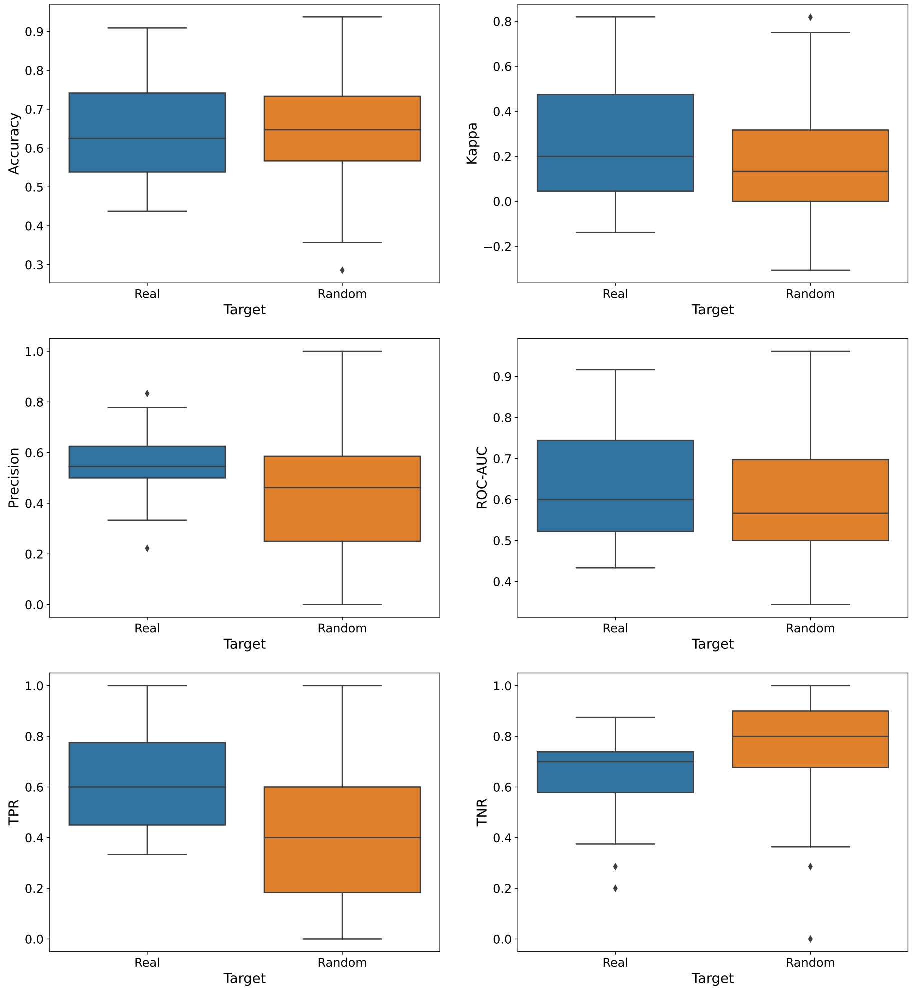


Figure B.18: Distributions of the different scores obtained for each prediction on the test set to check for overfitting in the VH valence. In dark blue, we have the model prediction through the actual target, while in orange, we find the model prediction of the overfitting check.

Bibliography

[1] R. W. Picard. Affective computing. Technical Report 321, Massachusetts Institute of Technology, 1995.

[2] Paul Ekman. *Basic Emotions*, chapter 3, pages 45–60. John Wiley Sons, Ltd, 1999.

[3] Sreeja PS and G Mahalakshmi. Emotion models: a review. *International Journal of Control Theory and Applications*, 10(8):651–657, 2017.

[4] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[5] Mimma Nardelli, Gaetano Valenza, Alberto Greco, Antonio Lanata, and Enzo Pasquale Scilingo. Recognizing emotions induced by affective sounds through heart rate variability. *IEEE Transactions on Affective Computing*, 6(4):385–394, 2015.

[6] Oliver Korn, Lukas Stamm, and Gerd Moeckl. Designing authentic emotions for non-human characters: A study evaluating virtual affective behavior. pages 477–487, 06 2017.

[7] Erik Cambria. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107, 2016.

[8] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.

[9] Stamos Katsigiannis and Naeem Ramzan. Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. *IEEE Journal of Biomedical and Health Informatics*, 22(1):98–107, 2018.

[10] Tim Dalgleish. The emotional brain. *Nature Reviews Neuroscience*, 5(7):583–589, 2004.

[11] Jorge L Armony, David Servan-Schreiber, Jonathan D Cohen, and Joseph E LeDoux. Computational modeling of emotion: Explorations through the anatomy and physiology of fear conditioning. *Trends in cognitive sciences*, 1(1):28–34, 1997.

[12] Maria Egger, Matthias Ley, and Sten Hanke. Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science*, 343:35–55, 2019. The proceedings of AmI, the 2018 European Conference on Ambient Intelligence.

[13] Arthur P Brief. *Attitudes in and around organizations*, volume 9. Sage, 1998.

[14] Herbert A Simon. *Administrative behavior*. Simon and Schuster, 2013.

[15] B Keith Payne. Prejudice and perception: the role of automatic and controlled processes in misperceiving a weapon. *Journal of personality and social psychology*, 81(2):181, 2001.

[16] Neal Schmitt. Method bias: The importance of theory and measurement. *Journal of Organizational Behavior*, pages 393–398, 1994.

[17] Sigal Barsade, Lakshmi Ramarajan, and Drew Westen. Implicit affect in organizations. *Research in Organizational Behavior - RES ORGAN BEH*, 29:135–162, 12 2009.

[18] Michael Brownstein, Alex Madva, and Bertram Gawronski. What do implicit measures measure? *Wiley interdisciplinary reviews. Cognitive science*, 10:e1501, 08 2019.

[19] Jan De Houwer and Agnes Moors. Implicit measures: Similarities and differences. *Handbook of implicit social cognition: Measurement, theory, and applications*, pages 176–193, 2010.

[20] Mariano Alcañiz, Elena Parra, and Isabela A. Chicchi Giglioli. Virtual reality as an emerging methodology for leadership assessment and training. *Frontiers in Psychology*, 9:1658, September 2018.

[21] Resham Arya, Jaiteg Singh, and Ashok Kumar. A survey of multidisciplinary domains contributing to affective computing. *Computer Science Review*, 40:100399, 2021.

[22] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. A review of emotion recognition using physiological signals. *Sensors*, 18(7), 2018.

[23] Andreas Riener, Alois Ferscha, and Mohamed Aly. Heart on the road: Hrv analysis for monitoring a driver's affective state. pages 99–106, 09 2009.

[24] Julian F Thayer, Fredrik Åhs, Mats Fredrikson, John J Sollers III, and Tor D Wager. A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews*, 36(2):747–756, 2012.

[25] A Craig, Y Tran, G Hermens, LM Williams, A Kemp, C Morris, and Evian Gordon. Psychological and neural correlates of emotional intelligence in a large sample of adult males and females. *Personality and Individual Differences*, 46(2):111–115, 2009.

[26] Xin Hu, Jingjing Chen, Fei Wang, and Dan Zhang. Ten challenges for eeg-based affective computing. *Brain Science Advances*, 5(1):1–20, 2019.

[27] Reiner Nikula. Psychological correlates of nonspecific skin conductance responses. *Psychophysiology*, 28(1):86–90, 1991.

[28] Henrique Sequeira, Pascal Hot, Laetitia Silvert, and Sylvain Delplanque. Electrical autonomic correlates of emotion. *International journal of psychophysiology*, 71(1):50–56, 2009.

[29] Cornelia Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, Gerhard Tröster, and Ulrike Ehlert. Discriminating stress from cognitive load using a wearable eda device. *IEEE Transactions on information technology in biomedicine*, 14(2):410–417, 2009.

[30] Andreas Glöckner and Ann-Katrin Herbold. An eye-tracking study on information processing in risky decisions: Evidence for compensatory strategies based on automatic processes. *Journal of Behavioral Decision Making*, 24(1):71–98, 2011.

[31] Hua Wang, Mark Chignell, and Mitsuru Ishizuka. Empathic tutoring software agents using real-time eye tracking. In *Proceedings of the 2006 symposium on Eye tracking research & applications*, pages 73–78, 2006.

[32] Günther Knoblich, Stellan Ohlsson, and Gary E Raney. An eye movement study of insight problem solving. *Memory & cognition*, 29(7):1000–1009, 2001.

[33] Kiavash Bahreini, Rob Nadolski, and Wim Westera. Towards multimodal emotion recognition in e-learning environments. *Interactive Learning Environments*, 24(3):590–605, 2016.

[34] Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Mingyue Niu, and Jiangyan Yi. Speech emotion recognition using semi-supervised learning with ladder networks. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pages 1–5. IEEE, 2018.

[35] Yu-Liang Hsu, Jeen-Shing Wang, Wei-Chun Chiang, and Chien-Han Hung. Automatic ecg-based emotion recognition in music listening. *IEEE Transactions on Affective Computing*, 11(1):85–99, 2020.

[36] Ian Daly, Asad Malik, James Weaver, Faustina Hwang, Slawmoir J Nasuto, Duncan Williams, Alexis Kirke, and Eduardo Miranda. Identifying music-induced emotions from eeg for use in brain-computer music interfacing. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 923–929. IEEE, 2015.

[37] Yixiang Dai, Xue Wang, Pengbo Zhang, and Weihang Zhang. Wearable biosensor network enabled multimodal daily-life emotion recognition employing reputation-driven imbalanced fuzzy classification. *Measurement*, 109:408–424, 2017.

[38] George Boateng. Towards real-time multimodal emotion recognition among couples. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, page 748–753, New York, NY, USA, 2020. Association for Computing Machinery.

[39] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953, 2019.

[40] Charles Clifton, Fernanda Ferreira, John M. Henderson, Albrecht W. Inhoff, Simon P. Liversedge, Erik D. Reichle, and Elizabeth R. Schotter. Eye movements in reading and information processing: Keith rayner's 40year legacy. *Journal of Memory and Language*, 86:1–19, 2016.

[41] Mariano Alcañiz, Irene Alice Chicchi-Giglioli, Lucía A. Carrasco-Ribelles, Javier Marín-Morales, Maria Eleonora Minissi, Gonzalo Teruel-García, Marian Sirera, and Luis Abad. Eye gaze as a biomarker in the recognition of autism spectrum disorder using virtual reality and machine learning: A proof of concept for diagnosis. *Autism Research*, 15(1):131–145, 2022.

[42] Hajra Ashraf, Mikael H Sodergren, Nabeel Merali, George Mylonas, Harsimrat Singh, and Ara Darzi. Eye-tracking technology in medical education: A systematic review. *Medical teacher*, 40(1):62–69, 2018.

[43] Sherif Said, Samer AlKork, T Beyrouthy, M Hassan, O Abdellatif, and M Fayek Abdraboo. Real time eye tracking and detection-a driving assistance system. *Advances in Science, Technology and Engineering Systems Journal*, 3(6):446–454, 2018.

[44] Frederic Martini and Edwin Bartholomew. *Essentials of Anatomy & Physiology*. Benjamin Cummings, San Francisco, California, 2001.

[45] Neil Carlson. *Physiology of Behavior*. Pearson Education, Inc., New York City, 2013.

[46] Richard Pflanzer. Galvanic skin response and the polygraph. PDF, 2014. Archived from the original (PDF) on 18 December 2014.

[47] Yun Liu and Siqing Du. Psychological stress level detection based on electrodermal activity. *Behavioural Brain Research*, 341:50–53, 2018.

[48] Mariano Alcañiz Raya, Irene Alice Chicchi Giglioli, Javier Marín-Morales, Juan L Higuera-Trujillo, Elena Olmos, Maria E Minissi, Gonzalo Teruel Garcia, Marian

Sirera, and Luis Abad. Application of supervised machine learning for behavioral biomarkers of autism spectrum disorder based on electrodermal activity and virtual reality. *Frontiers in human neuroscience*, 14:90, 2020.

[49] Francesco Chiossi, Robin Welsch, Steeven Villa, Lewis Chuang, and Sven Mayer. Virtual reality adaptation using electrodermal activity to support the user experience. *Big Data and Cognitive Computing*, 6(2), 2022.

[50] P. Sharma et al. Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. In A. Reis, J. Barroso, P. Martins, A. Jimoyiannis, R. YM. Huang, and R. Henriques, editors, *Technology and Innovation in Learning, Teaching and Education*, volume 1720 of *Communications in Computer and Information Science*. Springer, 2022.

[51] Luma Tabbaa, Ryan Searle, Saber Mirzaee Bafti, Md Moinul Hossain, Jittrapol Intarasisrisawat, Maxine Glancy, and Chee Siang Ang. Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 5(4):1–20, 2021.

[52] Raimondas Zemblys, Diederick C Niehorster, Oleg Komogortsev, and Kenneth Holmqvist. Using machine learning to detect events in eye-tracking data. *Behavior Research Methods*, 50(1):160–181, Feb 2018. Erratum in: Behavior Research Methods. 2019 Feb;51(1):451-452.

[53] Paulo Augusto de Lima Medeiros, Gabriel Vinícius Souza da Silva, Felipe Ricardo dos Santos Fernandes, Ignacio Sánchez-Gendriz, Hertz Wilton Castro Lins, Daniele Montenegro da Silva Barros, Danilo Alves Pinto Nagem, and Ricardo Alexsandro de Medeiros Valentim. Efficient machine learning approach for volunteer eye-blink detection in real-time using webcam. *Expert Systems with Applications*, 188:116073, 2022.

[54] Albert Bandura. The self system in reciprocal determinism. *American psychologist*, 33(4):344, 1978.

[55] Walter Mischel and Yuichi Shoda. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, 102(2):246, 1995.

[56] William Fleeson. Moving personality beyond the person-situation debate: The challenge and the opportunity of within-person variability. *Current Directions in Psychological Science*, 13(2):83–87, 2004.

[57] Bertram Gawronski and Galen V Bodenhausen. Evaluative conditioning from the perspective of the associative-propositional evaluation model. *Social Psychological Bulletin*, 13(3):1–33, 2018.

[58] Jacqueline M. Kory and Sidney K. D'Mello. Affect elicitation for affective computing. pages 371–383, 2015.

[59] Eddie Harmon-Jones and Cindy Harmon-Jones. Cognitive dissonance theory after 50 years of development. *Zeitschrift für Sozialpsychologie*, 38(1):7–16, 2007.

[60] Nicole A Roberts, Jeanne L Tsai, and James A Coan. Emotion elicitation using dyadic interaction tasks. *Handbook of emotion elicitation and assessment*, pages 106–123, 2007.

[61] Gaetano Valenza, Antonio lanatà, and Enzo Scilingo. Improving emotion recognition systems by embedding cardiorespiratory coupling. *Physiological measurement*, 34:449–464, 03 2013.

[62] Mohammad Soleymani, Maja Pantic, and Thierry Pun. Multimodal emotion recognition in response to videos (extended abstract). pages 491–497, 2015.

[63] Rosa M. Baños, Cristina Botella, Mariano Alcañiz, Víctor Liaño, Belén Guerrero, and Beatriz Rey. Immersion and emotion: Their impact on the sense of presence. *Cyberpsychology Behavior*, 7(6):734–741, December 2004.

[64] Ivan E. Sutherland. The ultimate display. 1965.

[65] Ivan E. Sutherland. A head-mounted three dimensional display. page 757–764, 1968.

[66] Tomasz Mazuryk and Michael Gervautz. Virtual reality - history, applications, technology and future. 12 1996.

[67] Sangsu Choi, Kiwook Jung, and Sang Do Noh. Virtual reality applications in manufacturing industries: Past research, present findings, and future directions. *Concurrent Engineering*, 23, 03 2015.

[68] Michael Zyda. From visual simulation to virtual reality to games. *Computer*, 38(9):25–32, 2005.

[69] Dara Meldrum, Aine Glennon, Susan Herdman, Deirdre Murray, and Rory McConn-Walsh. Virtual reality rehabilitation of balance: assessment of the usability of the nintendo wii® fit plus. *Disability and rehabilitation: assistive technology*, 7(3):205–210, 2012.

[70] Yang Song, Richard Koeck, and Shan Luo. Review and analysis of augmented reality (ar) literature for digital fabrication in architecture. *Automation in construction*, 128:103762, 2021.

[71] Claire Englund. Exploring approaches to teaching in three-dimensional virtual worlds. *The International Journal of Information and Learning Technology*, 34(2):140–151, 2017.

[72] Marianne Schmid Mast, Emmanuelle P Kleinlogel, Benjamin Tur, and Manuel Bachmann. The future of interpersonal skills development: Immersive virtual reality training with virtual humans. *Human Resource Development Quarterly*, 29(2):125–141, 2018.

[73] Anthony G Gallagher, E Matt Ritter, Howard Champion, Gerald Higgins, Marvin P Fried, Gerald Moses, C Daniel Smith, and Richard M Satava. Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. *Annals of surgery*, 241(2):364, 2005.

[74] Irene Alice Chicchi Giglioli, Gabriella Pravettoni, Dolores Lucia Sutil Martín, Elena Parra, and Mariano Luis Alcañiz Raya. A novel integrating virtual reality approach for the assessment of the attachment behavioral system. *Frontiers in Psychology*, 8, 2017.

[75] Javier Marín-Morales, Carmen Torrecilla-Moreno, Jaime Guixeres Provinciale, and María Del Carmen Llinares Millán. Methodological bases for a new platform for the measurement of human behaviour in virtual environments. *DYNA: Ingeniería e Industria*, 92(1):34–38, 2017.

[76] John Vince. *Introduction to Virtual Reality.* 01 2004.

[77] M. A. Raya, R. Baños, C. Botella, and Beatriz Rey. The emma project: Emotions as a determinant of presence. *PsychNology J.*, 1:141–150, 2003.

[78] G. Riva, F. Mantovani, C. S. Capideville, A. Preziosa, F. Morganti, D. Villani, A. Gaggioli, C. Botella, and M. Alcañiz. Affective interactions using virtual reality: The link between presence and emotions. *CyberPsychology & Behavior*, 10(1):45–56, 2007.

[79] Rosa María Baños, Víctor Liaño, Cristina Botella, Mariano Alcañiz, Belén Guerrero, and Beatriz Rey. Changing induced moods via virtual reality. In *Persuasive Technology: First International Conference on Persuasive Technology for Human Well-Being, PERSUASIVE 2006, Eindhoven, The Netherlands, May 18-19, 2006. Proceedings 1*, pages 7–15. Springer, 2006.

[80] Rosa María Baños, Ernestina Etchemendy, Diana Castilla, Azucena García-Palacios, Soledad Quero, and Cristina Botella. Positive mood induction procedures for virtual environments designed for elderly people. *Interacting with Computers*, 24(3):131–138, 2012.

[81] Alessandra Gorini, José Luis Mosso, Dejanira Mosso, Erika Pineda, Norma Leticia Ruíz, Miriam Ramíez, José Luis Morales, and Giuseppe Riva. Emotional response to virtual reality exposure across different cultures: the role of the attribution process. *Cyberpsychology & behavior*, 12(6):699–705, 2009.

[82] Alessandra Gorini, Claret S Capideville, Gianluca De Leo, Fabrizia Mantovani, and Giuseppe Riva. The role of immersion and narrative in mediated presence: the virtual hospital experience. *Cyberpsychology, behavior, and social networking*, 14(3):99–105, 2011.

[83] Alice Chirico, Pietro Cipresso, David B Yaden, Federica Biassoni, Giuseppe Riva, and Andrea Gaggioli. Effectiveness of immersive videos in inducing awe: an experimental study. *Scientific reports*, 7(1):1218, 2017.

[84] Jim Blascovich, Jack Loomis, Andrew C Beall, Kimberly R Swinth, Crystal L Hoyt, and Jeremy N Bailenson. Immersive virtual environment technology as a methodological tool for social psychology. *Psychological inquiry*, 13(2):103–124, 2002.

[85] Daniel Freeman, Sarah Reeve, Abi Robinson, Anke Ehlers, David Clark, Bernhard Spanlang, and Mel Slater. Virtual reality in the assessment, understanding, and treatment of mental health disorders. *Psychological medicine*, 47(14):2393–2400, 2017.

[86] Carla de Juan-Ripoll, José Llanes-Jurado, Irene Alice Chicchi Giglioli, Javier Marín-Morales, and Mariano Alcañiz. An immersive virtual reality game for predicting risk taking through the use of implicit measures. *Applied Sciences*, 11(2), 2021.

[87] Mel Slater and Maria V Sanchez-Vives. Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, 3:74, 2016.

[88] Cristina Botella, Javier Fernández-Álvarez, Verónica Guillén, Azucena García-Palacios, and Rosa Baños. Recent progress in virtual reality exposure therapy for phobias: a systematic review. *Current psychiatry reports*, 19:1–13, 2017.

[89] Roberto Lloréns, Enrique Noé, Carolina Colomer, and Mariano Alcañiz. Effectiveness, usability, and cost-benefit of a virtual reality–based telerehabilitation program for balance recovery after stroke: A randomized controlled trial. *Archives of Physical Medicine and Rehabilitation*, 96(3):418–425, 2015.

[90] Adrian Borrego, Jorge Latorre, Roberto Llorens, Mariano Alcañiz Raya, and Enrique Noé. Feasibility of a walking virtual reality system for rehabilitation: Objective and subjective parameters. *Journal of NeuroEngineering and Rehabilitation*, 13, 08 2016.

[91] Colin Ware, Kevin Arthur, and Kellogg S Booth. Fish tank virtual reality. pages 37–42, 1993.

[92] Matthew Lombard and Theresa Ditton. At the Heart of It All: The Concept of Presence. *Journal of Computer-Mediated Communication*, 3(2):JCMC321, 09 1997.

[93] J. M. Loomis, J. J. Blascovich, and A. C. Beall. Immersive virtual environment technology as a basic research tool in psychology. *Behavior Research Methods, Instruments, & Computers*, 31:557–564, 1999.

[94] Carrie Heeter. Interactivity in the context of designed experiences. *Journal of Interactive Advertising*, 1(1):3–14, 2000.

[95] Frank Biocca, Judee Burgoon, Chad Harms, Matt Stoner, et al. Criteria and scope conditions for a theory and measure of social presence. *Presence: Teleoperators and virtual environments*, 10(01):2001, 2001.

[96] Giovanni Vecchiato, Andrea Jelic, Gaetano Tieri, Anton Giulio Maglione, Francesco De Matteis, and Fabio Babiloni. Neurophysiological correlates of embodiment and motivational factors during the perception of virtual architectural environments. *Cognitive Processing*, 16:425–429, 2015.

[97] Lasse Jensen and Fredrik Konradsen. A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies*, 11:1–15, 2017.

[98] David Burden and Maggi Savin-Baden. *Virtual humans: Today and tomorrow.* CRC Press, 2019.

[99] Alaa A. Abd-alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M. Bewick, Peter Gardner, and Mowafa Househ. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, 132:103978, 2019.

[100] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45, jan 1966.

[101] V. Cerf. Parry encounters the doctor. RFC 439, January 1973.

[102] Jason L. Hutchens and Michael D. Alder. Introducing megahal. *CoNLL*, 1998.

[103] Samuel R. Bowman. Eight things to know about large language models. 2023.

[104] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

[105] A. M. TURING. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236):433–460, 10 1950.

[106] Philip Hingston. A turing test for computer game bots. *Computational Intelligence and AI in Games, IEEE Transactions on*, 1:169 – 186, 10 2009.

[107] N. Alvarado, S.S. Adams, S. Burbeck, and C. Latta. Beyond the turing test: performance metrics for evaluating a computer simulation of the human mind. In *Proceedings 2nd International Conference on Development and Learning. ICDL 2002*, pages 147–152, 2002.

[108] Xiaochuan Pan and Antonia F. C. Hamilton. Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology*, 109(3):395–417, August 2018.

[109] Salsa lipsync suite.

[110] Effie Karuzaki, Nikolaos Partarakis, Nikolaos Patsiouras, Emmanouil Zidianakis, Antonios Katzourakis, Antreas Pattakos, Danae Kaplanidi, Evangelia Baka, Nedjma Cadi, Nadia Magnenat-Thalmann, Chris Ringas, Eleana Tasiopoulou, and Xenophon Zabulis. Realistic virtual humans for cultural heritage applications. *Heritage*, 4(4):4148–4171, 2021.

[111] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.

[112] Epic Games. *Unreal Engine*, 2023. Accessed: November 5, 2023.

[113] FACEGOOD. FACEGOOD-Audio2Face github repository. `https://github.com/FACEGOOD/FACEGOOD-Audio2Face`, Accessed: November 10, 2023.

[114] Nina Döllinger, Erik Wolf, Mario Botsch, Marc Erich Latoschik, and Carolin Wienrich. Are embodied avatars harmful to our self-experience? the impact of virtual embodiment on body awareness. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2023.

[115] Eunjin (Anna) Kim, Donggyu Kim, Zihang E, and Heather Shoenberger. The next hype in social media advertising: Examining virtual influencers' brand endorsement effectiveness. *Frontiers in Psychology*, 14, 2023.

[116] Arturo S Garcia, Patricia Fernandez-Sotos, Miguel A Vicente-Querol, Guillermo Lahera, Roberto Rodriguez-Jimenez, and Antonio Fernandez-Caballero. Design of reliable virtual human facial expressions and validation by healthy people. *Integrated Computer-Aided Engineering*, 27(3):287–299, 2020.

[117] Effie Karuzaki, Nikolaos Partarakis, Nikolaos Patsiouras, Emmanouil Zidianakis, Antonios Katzourakis, Antreas Pattakos, Danae Kaplanidi, Evangelia Baka, Nedjma Cadi, Nadia Magnenat-Thalmann, Chris Ringas, Eleana Tasiopoulou, and Xenophon Zabulis. Realistic virtual humans for cultural heritage applications. *Heritage*, 4(4):4148–4171, 2021.

[118] Javier Marín-Morales, Carmen Llinares, Jaime Guixeres, and Mariano Alcañiz. Emotion recognition in immersive virtual reality: From statistics to affective computing. *Sensors*, 20(18), 2020.

[119] Jaime Guixeres, José Saiz, Mariano Alcañiz, Ausiàs Cebolla, Patricia Escobar, Rosa Baños, Cristina Botella, Juan F. Lison, Jose Alvarez, Lidia Cantero, and Empar Lurbe. Effects of virtual reality during exercise in children. *Journal of Universal Computer Science*, 19(9):1199–1218, 2013.

[120] Anna Felnhofer, Oswald D. Kothgassner, Mathias Schmidt, Anna K. Heinzle, Leon Beutl, Helmut Hlavacs, and Ilse Kryspin-Exner. Is virtual reality emotionally arousing? investigating five emotion-inducing virtual park scenarios. *International Journal of Human Computer Studies*, 82:48–56, 2015.

[121] Gonzalo Lorenzo, Antonio Lledó, Jorge Pomares, and Ruth Roig. Design and application of an immersive virtual reality system to enhance emotional skills for children with autism spectrum disorders. *Computers & Education*, 98:192–205, 2016.

[122] Javier Marín-Morales, Juan Luis Higuera Trujillo, Alberto Greco, Jaime Guixeres, Carmen Llinares, Enzo Scilingo, Mariano Alcañiz Raya, and Gaetano Valenza. Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors. *Scientific Reports*, 8, 09 2018.

[123] Vishnunarayan G Prabhu, Courtney Linder, Laura M Stanley, and Robert Morgan. An affective computing in virtual reality environments for managing surgical pain and anxiety. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 235–2351, 2019.

[124] Norma Ontiveros-Hernández, Miguel Pérez-Ramírez, and Yasmin Hernandez. Virtual reality and affective computing for improving learning. *Research in Computing Science*, 65:121–131, 12 2013.

[125] Depression. Retrieved 7 April 2021. Archived from the original on 26 December 2020.

[126] Peter De Zwart, Bertus Jeronimus, and Peter De Jonge. Empirical evidence for definitions of episode, remission, recovery, relapse and recurrence in depression: A systematic review. *Epidemiology and Psychiatric Sciences*, 28(5):544–562, 2019.

[127] P. Gilbert. *Psychotherapy and Counselling for Depression.* Sage, Los Angeles, 3rd edition, 2007.

[128] Pietro Cipresso, Irene Chicchi Giglioli, Mariano Alcañiz Raya, and Giuseppe Riva. The past, present, and future of virtual and augmented reality research: A network and cluster analysis of the literature. *Frontiers in Psychology*, 9:2086, 11 2018.

[129] J. Marín-Morales, J.L. Higuera-Trujillo, A. Greco, J. Guixeres, C. Llinares, C. Gentili, E.P. Scilingo, M. Alcañiz, and G. Valenza. Real vs. immersive-virtual emotional experience: Analysis of psycho-physiological patterns in a free exploration of an art museum. *PLOS ONE*, 14(10):1–24, 10 2019.

[130] Silvia Erika Kober, Jürgen Kurzmann, and Christa Neuper. Cortical correlate of spatial presence in 2d and 3d interactive virtual reality: An eeg study. *International Journal of Psychophysiology*, 83(3):365–374, 2012.

[131] Miriam Clemente, Alejandro Rodríguez, Beatriz Rey, and Mariano Alcañiz Raya. Assessment of the influence of navigation control and screen size on the sense of presence in virtual reality using eeg. *Expert Systems with Applications*, 41:1584–1592, 03 2014.

[132] Adrian Borrego, Jorge Latorre, Mariano Alcañiz Raya, and Roberto Llorens. Comparison of oculus rift and htc vive: Feasibility for virtual reality-based exploration, navigation, exergaming, and rehabilitation. *Games for Health Journal*, 7, 01 2018.

[133] Lasse Jensen and Flemming Konradsen. A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies*, 23:1–15, 07 2018.

[134] Tyler Jost, Grant Drewelow, Scott Koziol, and Jonathan Rylander. A quantitative method for evaluation of 6 degree of freedom virtual reality systems. *Journal of Biomechanics*, 97:109379, 10 2019.

[135] Tilanka Chandrasekera, Kinkini Fernando, and Luis Puig. Effect of degrees of freedom on the sense of presence generated by virtual reality (vr) head-mounted display systems: A case study on the use of vr in early design studios. *Journal of Educational Technology Systems*, 47:004723951882486, 01 2019.

[136] Oana Bălan, Gabriela Moise, Alin Moldoveanu, Marius Leordeanu, and Florica Moldoveanu. An investigation of various machine and deep learning techniques applied in automatic fear level detection and acrophobia virtual therapy. *Sensors*, 20(2), 2020.

[137] Thomas Armstrong and Bunmi Olatunji. Eye tracking of attention in the affective disorders: A meta-analytic review and synthesis. *Clinical psychology review*, 32:704–23, 09 2012.

[138] David E. Irwin. Memory for position and identity across eye movements. *Journal of Experimental Psychology: Learning Memory and Cognition*, 18(2):307–317, March 1992.

[139] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124 3:372–422, 1998.

[140] Vildan Tanriverdi and Robert Jacob. Interacting with eye movements in virtual environments. *Conference on Human Factors in Computing Systems - Proceedings*, 08 2001.

[141] Alexander Skulmowski, Andreas Bunge, Kai Kaspar, and Gordon Pipa. Forced-choice decision-making in modified trolley dilemma situations: A virtual reality and eye tracking study. *Frontiers in behavioral neuroscience*, 8:426, 12 2014.

[142] Joshua Juvrud, Gustaf Gredebäck, Fredrik Åhs, Nils Lerin, Pär Nyström, Granit Kastrati, and Jörgen Rosén. The immersive virtual reality lab: Possibilities for remote experimental manipulations of autonomic activity on a large scale. *Frontiers in Neuroscience*, 12, 04 2018.

[143] Viviane Clay, Peter König, and Sabine Koenig. Eye tracking in virtual reality. *Journal of Eye Movement Research*, 12, 04 2019.

[144] Roy Hessels, Diederick Niehorster, Marcus Nyström, Richard Andersson, and Ignace Hooge. Is the eye-movement field confused about fixations and saccades? a survey among 124 researchers. *Royal Society Open Science*, 5:180502, 08 2018.

[145] Gabriel Diaz, Joseph Cooper, Dmitry Kit, and Mary Hayhoe. Real-time recording and classification of eye movements in an immersive virtual environment. *Journal of vision*, 13, 10 2013.

[146] Otto Lappi. Eye tracking in the wild: the good, the bad and the ugly. *Journal of Eye Movement Research*, 8:1, 10 2015.

[147] Andrew Duchowski, Eric Medlin, Anand Gramopadhye, Brian Melloy, and Santosh Nair. Binocular eye tracking in vr for visual inspection training. pages 1–8, 11 2001.

[148] Jia Zheng Lim, James Mountstephens, and Jason Teo. Emotion recognition using eye-tracking: Taxonomy, review and current challenges. *Sensors*, 20(8), 2020.

[149] Barry Manor and Evian Gordon. Defining the temporal threshold for ocular fixation in free-viewing visuocogitive tasks. *Journal of neuroscience methods*, 128:85–93, 09 2003.

[150] Dario Salvucci and Joseph Goldberg. Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the Eye Tracking Research and Applications Symposium*, pages 71–78, 01 2000.

[151] Andrew Duchowski, Eric Medlin, Nathan Cournia, Hunter Murphy, Anand Gramopadhye, Santosh Nair, Jeenal Vorah, and Brian Melloy. 3-d eye movement analysis. *Behavior research methods, instruments, computers : a journal of the Psychonomic Society, Inc*, 34:573–91, 12 2002.

[152] Andrew Duchowski. *Eye Tracking Methodology: Theory and Practice*. 01 2007.

[153] Vladislava Bobić and Stevica Graovac. Development, implementation and evaluation of new eye tracking methodology. pages 1–4, 2016.

[154] Najood Al Ghamdi and Wadee Alhalabi. Fixation detection with ray-casting in immersive virtual reality. *International Journal of Advanced Computer Science and Applications*, 10, 01 2019.

[155] Ludwig Sidenmark and Anders Lundström. Gaze behaviour on interacted objects during hand interaction in virtual reality for eye tracking calibration. pages 1–9, 06 2019.

[156] Frederick Shic, Brian Scassellati, and Katarzyna Chawarska. The incomplete fixation measure. pages 111–114, 01 2008.

[157] Pieter Blignaut. Fixation identification: The optimum threshold for a dispersion algorithm. *Attention, perception psychophysics*, 71:881–95, 06 2009.

[158] W. Boucsein. *Electrodermal activity.* Springer Science+Business Media, LLC, New York, 2012.

[159] PH Ellaway, A Kuppuswamy, A Nicotra, and CJ Mathias. Sweat production and the sympathetic skin response: improving the clinical assessment of autonomic function. *Autonomic neuroscience : basic amp; clinical*, 155(1-2):109—114, June 2010.

[160] Mathias Benedek and Christian Kaernbach. A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, 190(1):80 – 91, 2010.

[161] Hugo Posada-Quintero and Kaye Chon. Innovations in electrodermal activity data collection and signal processing: A systematic review. *Sensors*, 20:479, 01 2020.

[162] Michael Dawson, Anne Schell, and Diane Filion. The electrodermal system. 01 2000.

[163] Alberto Greco, Gaetano Valenza, and Enzo Scilingo. *Advances in Electrodermal Activity Processing with Applications for Mental Health.* 01 2016.

[164] A S Anusha, Joy Jose, S P Preejith, Joseph Jayaraj, and Sivaprakasam Mohanasankar. Physiological signal based work stress detection using unobtrusive sensors. *Biomedical Physics & Engineering Express*, 4(6):065001, sep 2018.

[165] Roberto Zangróniz, Arturo Martínez Rodrigo, José Manuel Pastor García, María López Bonal, and Antonio Fernández-Caballero. Electrodermal activity sensor for classification of calm/distress condition. *Sensors*, 17:2324, 10 2017.

[166] Yun Liu and Siqing Du. Psychological stress level detection based on electrodermal activity. *Behavioural Brain Research*, 341:50–53, 2018.

[167] A. Greco, G. Valenza, A. Lanatà, G. Rota, and E. Scilingo. Electrodermal activity in bipolar patients during affective elicitation. *IEEE journal of biomedical and health informatics*, 18:1865–1873, 11 2014.

[168] Giulia Perugia, Daniel Rodriguez-Martin, Marta DIaz Boladeras, Andreu Catala Mallofre, Emilia Barakova, and Matthias Rauterberg. In *RO-MAN 2017 - 26th IEEE International Symposium on Robot and Human Interactive Communication*, pages 1248–1254, United States, December 2017. Institute of Electrical and Electronics Engineers. 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2017), RO-MAN 2017 ; Conference date: 28-08-2017 Through 01-09-2017.

[169] Jainendra Shukla, Miguel Barreda-Ángeles, Joan Oliver, and Domènec Puig. Efficient wavelet-based artifact removal for electrodermal activity in real-world applications. *Biomedical Signal Processing and Control*, 42:45 – 52, 2018.

[170] D Malathi, JD Dorathi Jayaseeli, S Madhuri, and K Senthilkumar. Electrodermal activity based wearable device for drowsy drivers. *Journal of Physics: Conference Series*, 1000:012048, apr 2018.

[171] Iolanda Leite, Rui Henriques, Carlos Martinho, and Ana Paiva. Sensors in the wild: Exploring electrodermal activity in child-robot interaction. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '13, page 41–48. IEEE Press, 2013.

[172] Jeongmi (Jamie) Kim and Daniel R. Fesenmaier. Measuring emotions in real time: Implications for tourism experience design. *Journal of Travel Research*, 54(4):419–429, 2015.

[173] Irene Chicchi Giglioli, Gabriella Pravettoni, Lucia Sutil, Elena Parra, and Mariano Alcañiz Raya. A novel integrating virtual reality approach for the assessment of the attachment behavioral system. *Frontiers in Psychology*, 8, 06 2017.

[174] Jeff Tarrant, Jeremy Viczko, and Hannah Cope. Virtual reality for anxiety reduction demonstrated by quantitative eeg: A pilot study. *Frontiers in Psychology*, 9:1280, 2018.

[175] Valentina Matijević, Ana Šečić, Valentina Mašić, Martina Sunić, Zeljka Kolak, and Mateja Znika. Virtual reality in rehabilitation and therapy. *Acta clinica Croatica*, 52:453–7, 12 2013.

[176] A. Li, Z. Montaño, V. J. Chen, and J. Gold. Virtual reality and pain management: current trends and future directions. *Pain management*, 1 2:147–157, 2011.

[177] E. Bekele, D. Bian, J. Peterman, S. Park, and N. Sarkar. Design of a virtual reality system for affect analysis in facial expressions (vr-saafe); application to schizophrenia. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(6):739–749, 2017.

[178] Rytis Maskeliunas, Justas Šalkevicius, Robertas Damaševičius, Rytis Maskeliunas, and Ilona Laukienė. Anxiety level recognition for virtual reality therapy system using physiological signals. *Electronics*, 8(9):1039, 2019.

[179] J. Kritikos, G. Tzannetos, C. Zoitaki, S. Poulopoulou, and D. Koutsouris. Anxiety detection from electrodermal activity sensor with movement interaction during virtual reality simulation. In *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 571–576, 2019.

[180] Débora Salgado, Felipe Martins, Thiago Braga Rodrigues, Conor Keighrey, Ronan Flynn, Eduardo Naves, and Niall Murray. A qoe assessment method based on eda, heart rate and eeg of a virtual reality assistive technology system. pages 517–520, 06 2018.

[181] Mariano Alcañiz Raya, Irene Alice Chicchi Giglioli, Javier Marín-Morales, Juan L. Higuera-Trujillo, Elena Olmos, Maria E. Minissi, Gonzalo Teruel Garcia, Marian Sirera, and Luis Abad. Application of supervised machine learning for behavioral biomarkers of autism spectrum disorder based on electrodermal activity and virtual reality. *Frontiers in Human Neuroscience*, 14:90, 2020.

[182] S. Taylor, N. Jaques, W. Chen, S. Fedor, A. Sano, and R. Picard. Automatic identification of artifacts in electrodermal activity data. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1934–1937, 2015.

[183] Dominik R. Bach. A head-to-head comparison of scralyze and ledalab, two model-based methods for skin conductance analysis. *Biological Psychology*, 103:63 – 68, 2014.

[184] Weixuan Chen, Natasha Jaques, Sara Taylor, Akane Sano, Szymon Fedor, and Rosalind W. Picard. Wavelet-based motion artifact removal for electrodermal activity. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6223–6226, 2015.

[185] Yuning Zhang, Maysam Haghdan, and Kevin S. Xu. Unsupervised motion artifact detection in wrist-measured electrodermal activity data. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, ISWC '17, page 54–57, New York, NY, USA, 2017. Association for Computing Machinery.

[186] Md-Billal Hossain, Hugo F Posada-Quintero, Youngsun Kong, Riley McNaboe, and Ki H Chon. Automatic motion artifact detection in electrodermal activity data using machine learning. *Biomedical Signal Processing and Control*, 74:103483, 2022.

[187] Ian R Kleckner, Rebecca M Jones, Oliver Wilder-Smith, Jolie B Wormwood, Murat Akcakaya, Karen S Quigley, Catherine Lord, and Matthew S Goodwin. Simple, transparent, and flexible automated quality assessment procedures for ambulatory electrodermal activity data. *IEEE transactions on bio-medical engineering*, 65(7):1460—1467, July 2018.

[188] Shkurta Gashi, Elena Di Lascio, Bianca Stancu, Vedant Das Swain, Varun Mishra, Martin Gjoreski, and Silvia Santini. Detection of artifacts in ambulatory electrodermal activity data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(2), June 2020.

[189] Sandya Subramanian, Bryan Tseng, Riccardo Barbieri, and Emery N Brown. An unsupervised automated paradigm for artifact removal from electrodermal activity in an uncontrolled clinical setting. *Physiological Measurement*, 43(11):115005, 2022.

[190] Javier Hernandez, Rob R. Morris, and Rosalind W. Picard. Call center stress recognition with person-specific models. In Sidney D'Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin, editors, *Affective Computing and Intelligent Interaction*, pages 125–134, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[191] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. cvxeda: A convex optimization approach to electrodermal activity processing. *IEEE Trans. Biomed. Eng.*, 63(4):797–804, 2016.

[192] Nagarajan Ganapathy, Yedukondala Rao Veeranki, and Ramakrishnan Swaminathan. Convolutional neural network based emotion classification using electro-

dermal activity signals and time-frequency features. *Expert Systems with Applications*, 159:113571, 2020.

[193] Yekta Can, Niaz Chalabianloo, Deniz Ekiz, and Cem Ersoy. Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study. *Sensors*, 19, 04 2019.

[194] Md Billal Hossain, Hugo Posada-Quintero, and Ki Chon. A deep convolutional autoencoder for automatic motion artifact removal in electrodermal activity. *IEEE Transactions on Biomedical Engineering*, 2022.

[195] Shaveta Dargan, Munish Kumar, · Maruthi, Maruthi Rohit Ayyagari, and · Kumar. A survey of deep learning and its applications: A new paradigm to machine learning. 07 2019.

[196] Monika Bansal, Munish Kumar, and Manish Kumar. 2d object recognition techniques: State-of-the-art work. *Archives of Computational Methods in Engineering*, 28:1147–1161, 2020.

[197] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[198] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[199] Sreenath P. Kyathanahally, André Döring, and Roland Kreis. Deep learning approaches for detection and removal of ghosting artifacts in mr spectroscopy. *Magnetic Resonance in Medicine*, 80(3):851–863, 2018.

[200] Nuno Bento, David Belo, and Hugo Gamboa. Ecg biometrics using spectrograms and deep neural networks. 02 2020.

[201] Manhua Liu, Fan Li, Hao Yan, Kundong Wang, Yixin Ma, Li Shen, and Mingqing Xu. A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in alzheimer's disease. *NeuroImage*, 208:116459, 2020.

[202] Karol Antczak. Deep recurrent neural networks for ECG signal denoising. *CoRR*, abs/1807.11551, 2018.

[203] Jose Llanes-Jurado, Lucia Carrasco-Ribelles, Mariano Alcañiz, and Javier Marín-Morales. Electrodermal activity artifact correction benchmark (edabe), January 2023.

[204] Adi Hajj-Ahmad, Ravi Garg, and Min Wu. Enf-based region-of-recording identification for media signals. *IEEE Transactions on Information Forensics and Security*, 10(6):1125–1136, 2015.

[205] Jihye Moon, Md Billal Hossain, and Ki H. Chon. Ar and arma model order selection for time-series modeling with imagenet classification. *Signal Processing*, 183:108026, 2021.

[206] Hengliang Wang, Kin Siu, Kihwan Ju, and hc ki. A high resolution approach to estimating time-frequency spectra and their amplitudes. *Annals of biomedical engineering*, 34:326–38, 03 2006.

[207] Swarnendu Ghosh, Nibaran Das, Ishita Das, and Ujjwal Maulik. Understanding deep learning techniques for image segmentation. *ACM Comput. Surv.*, 52(4), August 2019.

[208] Marieke van Dooren, Joris H Janssen, et al. Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiology & behavior*, 106(2):298–304, 2012.

[209] Rosalind W. Picard. Affective computing: challenges. *International Journal of Human-Computer Studies*, 59(1):55–64, 2003. Applications of Affective Computing in Human-Computer Interaction.

[210] Atefeh Safayari and Hamidreza Bolhasani. Depression diagnosis by deep learning using eeg signals: A systematic review, 07 2021.

[211] Sukit Suparatpinyo and Nuanwan Soonthornphisaj. Smart voice recognition based on deep learning for depression diagnosis. *Artificial Life and Robotics*, pages 1–11, 2023.

[212] Justin Brian Balano, Vanessa Ley Huerto, Sigfried Sanchez, Aresh Saharkhiz, and Joel De Goma. Determining the level of depression using bdi-ii through voice recognition. In *2019 IEEE 6th International Conference on Industrial Engineering and Applications (ICIEA)*, pages 387–391. IEEE, 2019.

[213] James J Gross and Hooria Jazaieri. Emotion, emotion regulation, and psychopathology: An affective science perspective. *Clinical psychological science*, 2(4):387–401, 2014.

[214] Shalom Greene, Himanshu Thapliyal, and Allison Caban-Holt. A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health. *IEEE Consumer Electronics Magazine*, 5(4):44–56, 2016.

[215] Christy Ludlow. Central nervous system control of voice and swallowing. *Journal of clinical neurophysiology : official publication of the American Electroencephalographic Society*, 32:294–303, 08 2015.

[216] Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, 21(4):1249, 2021.

[217] Margaret M. Bradley and Peter J. Lang. The international affective picture system (iaps) in the study of emotion and attention. 2007.

[218] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis ;using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[219] Sahinya Susindar, Mahnoosh Sadeghi, Lea Huntington, Andrew Singer, and Thomas Ferris. The feeling is real: Emotion elicitation in virtual reality. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63:252–256, 11 2019.

[220] D. Burden and M. Savin-Baden. *Virtual Humans: Today and Tomorrow*. Chapman & Hall / CRC Artificial intelligence and robotics series. CRC Press, 2019.

[221] Ibrahim F. Imam and Yves Kodratoff. Intelligent adaptive agents: A highlight of the field and the AAAI-96 workshop. *AI Mag.*, 18(3):75–80, 1997.

[222] Iulian Vlad Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron Courville. Multiresolution recurrent neural networks: An application to dialogue response generation, 2016.

[223] M Hachman. Battle of the digital assistants: Cortana, siri, and google now. *PC World*, 32(6):13, 2014.

[224] Edward Moemeka and Elizabeth Moemeka. Leveraging cortana and speech. In *Real world windows 10 development*, pages 471–520. Springer, 2015.

[225] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

[226] AnaÃ¯s Tack and Chris Piech. The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues. In Antonija Mitrovic and Nigel Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 522–529, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

[227] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[228] P Antonius Angga, W Edwin Fachri, A Elevanita, Suryadi, and R Dewi Agushinta. Design of chatbot with 3d avatar, voice interface, and facial expression. In *2015 International Conference on Science in Information Technology (ICSITech)*, pages 326–330, 2015.

[229] Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems*, 92:539–548, 2019.

[230] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jonathan Gratch, Arno Hartholt, Margot Lor-Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, and Louis-Philippe Morency.

Simsensei kiosk: A virtual human interviewer for healthcare decision support. *13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014*, 2:1061–1068, 01 2014.

[231] Teresa Sollfrank, Oona Kohnen, Peter Hilfiker, Lorena C. Kegel, Hennric Jokeit, Peter Brugger, Miriam L. Loertscher, Anton Rey, Dieter Mersch, Joerg Sternagel, Michel Weber, and Thomas Grunwald. The effects of dynamic and static emotional facial expressions of humans and their avatars on the eeg: An erp and erd/ers study. *Frontiers in Neuroscience*, 15, 2021.

[232] Natalie Hube, Kresimir Vidackovic, and Michael Sedlmair. Using expressive avatars to increase emotion recognition: A pilot study. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA, 2022. Association for Computing Machinery.

[233] David Galin, R Ornstein, J Herron, and J Johnstone. Sex and handedness differences in eeg measures of hemispheric specialization. *Brain and language*, 16(1):19–55, 1982.

[234] Alejandra Ospina-Bohórquez, Sara Rodríguez-González, and Diego Vergara-Rodríguez. A review on multi-agent systems and virtual reality. In *Distributed Computing and Artificial Intelligence, Volume 1: 18th International Conference 18*, pages 32–42. Springer, 2022.

[235] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[236] Hervé Bredin and Antoine Laurent. End-to-end speaker segmentation for overlap-aware resegmentation. 2021.

[237] P Lang. Behavioral treatment and bio-behavioral assessment: Computer applications. *Technology in mental health care delivery systems*, pages 119–137, 1980.

[238] M. M. Bradley and P Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[239] Alejandro Guillén-Riquelme and Gualberto Casal. Actualización psicométrica y funcionamiento diferencial de los ítems en el state trait anxiety inventory (stai). *Psicothema*, 23, 01 2011.

[240] Fred Miao, Irina V Kozlenkova, Haizhong Wang, Tao Xie, and Robert W Palmatier. An emerging theory of avatar marketing. *Journal of Marketing*, 86(1):67–90, 2022.

[241] Julia Diemer, Georg W Alpers, Henrik M Peperkorn, Youssef Shiban, and Andreas Mühlberger. The impact of perception and presence on emotional reactions: a review of research in virtual reality. *Frontiers in psychology*, 6:26, 2015.

[242] Albert Mehrabian and Susan R Ferris. Inference of attitudes from nonverbal communication in two channels. *Journal of consulting psychology*, 31(3):248, 1967.

[243] Amrisha Vaish, Tobias Grossmann, and Amanda Woodward. Not all emotions are created equal: the negativity bias in social-emotional development. *Psychological bulletin*, 134(3):383, 2008.

[244] Lorainne Tudor Car, Dhakshenya Ardhithy Dhinagaran, Bhone Myint Kyaw, Tobias Kowatsch, Shafiq Joty, Yin-Leng Theng, and Rifat Atun. Conversational agents in health care: scoping review and conceptual analysis. *Journal of medical Internet research*, 22(8):e17158, 2020.

[245] John Torous, Sandra Bucci, Imogen H Bell, Lars V Kessing, Maria Faurholt-Jepsen, Pauline Whelan, Andre F Carvalho, Matcheri Keshavan, Jake Linardon, and Joseph Firth. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*, 20(3):318–335, 2021.

[246] Achini Adikari, Daswin De Silva, Harsha Moraliyage, Damminda Alahakoon, Jiahui Wong, Mathew Gancarz, Suja Chackochan, Bomi Park, Rachel Heo, and Yvonne Leung. Empathic conversational agents for real-time monitoring and co-facilitation of patient-centered healthcare. *Future Generation Computer Systems*, 126:318–329, 2022.

[247] Richard May and Kerstin Denecke. Extending patient education with claire: an interactive virtual reality and voice user interface application. In *Addressing Global Challenges and Quality Education: 15th European Conference on Technology Enhanced Learning, EC-TEL 2020, Heidelberg, Germany, September 14–18, 2020, Proceedings 15*, pages 482–486. Springer, 2020.

[248] Ghazala Bilquise, Samar Ibrahim, Khaled Shaalan, et al. Emotionally intelligent chatbots: A systematic literature review. *Human Behavior and Emerging Technologies*, 2022, 2022.

[249] Kerstin Denecke and Richard May. Developing a technical-oriented taxonomy to define archetypes of conversational agents in health care: Literature review and cluster analysis. *Journal of Medical Internet Research*, 25:e41583, 2023.

[250] Iulia Stanica, Maria-Iuliana Dascalu, Constanta Nicoleta Bodea, and Alin Dragos Bogdan Moldoveanu. Vr job interview simulator: where virtual reality meets artificial intelligence for education. In *2018 Zooming innovation in consumer technologies conference (ZINC)*, pages 9–12. IEEE, 2018.

[251] Chinedu Wilfred Okonkwo and Abejide Ade-Ibijola. Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, 2:100033, 2021.

[252] Chris Segrin. Social skills deficits associated with depression. *Clinical psychology review*, 20(3):379–403, 2000.

[253] Wootaek Lim, Daeyoung Jang, and Taejin Lee. Speech emotion recognition using convolutional and recurrent neural networks. pages 1–4, 2016.

[254] Ali Ghofrani, Rahil Mahdian Toroghi, and Shirin Ghanbari. Realtime face-detection and emotion recognition using mtcnn and minishufflenet v2. pages 817–821, 2019.

[255] Cong Zong and Mohamed Chetouani. Hilbert-huang transform based physiological signals analysis for emotion recognition. pages 334–339, 2009.

[256] Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, and Remigiusz Rak. Eye-tracking analysis for emotion recognition. *Computational Intelligence and Neuroscience*, 2020:1–13, 09 2020.

[257] Houwei Cao and Forest Elliott. Analysis of eye fixations during emotion recognition in talking faces. pages 1–7, 2021.

[258] Claudio Aracena, Sebastián Basterrech, Vaclav Snasel, and Juan Velasquez. Neural networks for emotion recognition based on eye tracking data. pages 2632–2637, 10 2015.

[259] Chin Leong Lim, Charles Rennie, Robert J. Barry, Hooman Bahramali, Ilene Lazzaro, Barry Manor, and Evian Gordon. Decomposing skin conductance into tonic and phasic components. *International Journal of Psychophysiology*, 25(2):97–109, 1997.

[260] Mahima Sharma, Sudhanshu Kacker, and Mohit Sharma. A brief introduction and review on galvanic skin response. *International Journal of Medical Research Professionals*, 2, 12 2016.

[261] Deger Ayata, Yusuf Yaslan, and Mustafa Kamasak. Emotion recognition via random forest and galvanic skin response. pages 1–4, 10 2016.

[262] Praveen Govarthan, Sriram P, Nagarajan Ganapathy, and Jack Fredo. Deep learning framework for categorical emotional states assessment using electrodermal activity signals. *Studies in health technology and informatics*, 305:40–43, 06 2023.

[263] Jainendra Shukla, Miguel Barreda-Ángeles, Joan Oliver, G. C. Nandi, and Domènec Puig. Feature extraction and selection for emotion recognition from electrodermal activity. *IEEE Transactions on Affective Computing*, 12(4):857–869, 2021.

[264] Maura Mengoni, Michele Germani, and Margherita Peruzzini. Benchmarking of virtual reality performance in mechanics education. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 5:103–117, 06 2011.

[265] Mel Slater. Place illusion and plausibility can lead to realistic behavior in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3549–57, 2009.

[266] Joana Pinto, Ana Fred, and Hugo Plácido da Silva. Biosignal-based multimodal emotion recognition in a valence-arousal affective framework applied to immersive video visualization. pages 3577–3583, 2019.

[267] Miguel Vicente-Querol, Antonio Fernández-Caballero, Jose Pascual Molina Masso, Pascual González, Luz González-Gualda, Patricia Fernández-Sotos, and Arturo García. *Influence of the Level of Immersion in Emotion Recognition Using Virtual Humans*, pages 464–474. 05 2022.

[268] OpenAI. Gpt-4 technical report, 2023.

[269] Samantha L. Finkelstein, Andrea Nickel, Lane Harrison, Evan A. Suma, and Tiffany Barnes. cmotion: A new game design to teach emotion recognition and programming logic to children using virtual humans. In *2009 IEEE Virtual Reality Conference*, pages 249–250, 2009.

[270] Katja Zibrek, Ludovic Hoyet, Kerstin Ruhland, and Rachel McDonnell. Evaluating the effect of emotion on gender recognition in virtual humans. 08 2013.

[271] Funda Durupinar and Jiehyun Kim. Facial emotion recognition of virtual humans with different genders, races, and ages. In *ACM Symposium on Applied Perception 2022*, SAP '22, New York, NY, USA, 2022. Association for Computing Machinery.

[272] Jose Llanes-Jurado, Lucía Gómez-Zaragozá, Maria Eleonora Minissi, Mariano Alcañiz, and Javier Marín-Morales. Developing conversational virtual humans for social emotion elicitation based on large language models. *Expert Systems with Applications*, 246:123261, 2024.

[273] Saloni Dattani, Hannah Ritchie, and Max Roser. Mental health. *Our World in Data*, 2021. https://ourworldindata.org/mental-health.

[274] Richard D Wright and Lawrence M Ward. *Orienting of attention.* Oxford University Press, 2008.

[275] Thomas Suslow, Anja Husslack, Anette Kersting, and Charlott Maria Bodenschatz. Attentional biases to emotional information in clinical depression: A systematic and meta-analytic review of eye tracking findings. *Journal of Affective Disorders*, 274:632–642, 2020.

[276] Ah Kim, Eun-Hye Jang, Seunghwan Kim, Kwan Choi, Hong Jin Jeon, Han Yu, and Sangwon Byun. Automatic detection of major depressive disorder using electrodermal activity. *Scientific Reports*, 8, 11 2018.

[277] Xinfang Ding, Xinxin Yue, Rui Zheng, Cheng Bi, Dai Li, and Guizhong Yao. Classifying major depression patients and healthy controls using eeg, eye tracking and galvanic skin response data. *Journal of Affective Disorders*, 251:156–161, 2019.

[278] P. Philip, J. A. Micoulaud-Franchi, P. Sagaspe, E. Sevin, J. Olive, S. Bioulac, and A. Sauteraud. Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. *Sci Rep*, 7:42656, Feb 2017.

[279] Joy O. Egede, Dominic Price, Deepa B. Krishnan, Shashank Jaiswal, Natasha Elliott, Richard Morriss, Maria J. Galvez Trigo, Neil Nixon, Peter Liddle, Christopher Greenhalgh, and Michel Valstar. Design and evaluation of virtual human mediated tasks for assessment of depression and anxiety. page 52–59, 2021.

[280] A. Takemoto. Depression detection using virtual avatar communication and eye tracking system. *Journal of Eye Movement Research*, 16(2), 2023.

[281] Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. The phq-9. *Journal of General Internal Medicine*, 16(9):606–613, 2001.

[282] Jose Llanes-Jurado, Lucía A. Carrasco-Ribelles, Mariano Alcañiz, Emilio Soria-Olivas, and Javier Marín-Morales. Automatic artifact recognition and correction for electrodermal activity based on lstm-cnn models. *Expert Systems with Applications*, 230:120581, 2023.

[283] Bo Hjorth. Eeg analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, 29(3):306–310, 1970.

[284] L. F. Kozachenko and N. N. Leonenko. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

[285] B Scholkopf, Alex Smola, Robert Williamson, and Peter Bartlett. New support vector algorithms. *Neural computation*, 12:1207–45, 06 2000.

[286] Gilles Louppe. Understanding random forests: From theory to practice. 2015.