The Multimodal Turing Test for Realistic Humanoid Robots with Embodied Artificial Intelligence

Carl Strathearn [1] and Minhua Ma [2]

[1] School of Computing and Digital Technologies, Staffordshire University, UK. Carl.Strathearn@research.staffs.ac.uk
[2] Provost, Falmouth University, UK. M.Ma@falmouth.ac.uk

Abstract

Alan Turing developed the *Turing Test* as a method to determine whether artificial intelligence (AI) can deceive human interrogators into believing it is sentient by competently answering questions at a confidence rate of 30%+. However, the *Turing Test* is concerned with natural language processing (NLP) and neglects the significance of appearance, communication and movement. The theoretical proposition at the core of this paper: 'can machines emulate human beings?' is concerned with both functionality and materiality. Many scholars consider the creation of a realistic humanoid robot (RHR) that is perceptually indistinguishable from a human as the apex of humanity's technological capabilities. Nevertheless, no comprehensive development framework exists for engineers to achieve higher modes of human emulation, and no current evaluation method is nuanced enough to detect the causal effects of the Uncanny Valley (UV) effect. The Multimodal Turing Test (MTT) provides such a methodology and offers a foundation for creating higher levels of human likeness in RHRs for enhancing human-robot interaction (HRI)

Key Words: *Turing Test, Humanoid Robots, Artificial Intelligence, Embodied Artificial Intelligence, HRI*

1. Introduction

The Turing Test hypothetically evaluated computational AI using typesetters and pre-written scripture to emulate human thought (Turing, 1950). However, modern conversational AI systems function with greater accuracy at a higher rate of processing than the analogue methods outlined in Turing's paper. Landgrebe & Smith (2019) explain that unlike the original Turing Test, the updated Turing Test for AI utilises two computer interfaces to replace the type-setter methodology. One computer system implements a conversational AI application and the other controlled by a human agent concealed from the view of the human interrogator. The role of the human interrogator is to evaluate the authenticity and accuracy of the agent's responses to determine which system is artificial and which is human. There are accounts of AI systems which claim to have passed the Turing Test. For example, Warwick & Shah (2015) and Aamoth (2014), advocate that a chatbot program named Eugene Goostman passed the 30% benchmark of the Turing Test in 2014, scoring a marginal 33%, at the Royal Society AI competition in 2014. However, commentators such as Copeland (2014), Hern (2014) and Robbins (2014) contest the validity of this achievement, stating two significant flaws in the evaluation procedure. Firstly, human interrogators had prior knowledge that the AI system emulated a 13-year old Ukrainian boy. This approach dissolves the integrity of the Turing Test, which states the removal of all identifiers is vital in maintaining impartiality (Turing, 1950). Secondly, the creators of the Eugene Goostman chatbot hand-selected the human interrogators for the test, significantly increasing the probability for participant bias. Sample & Hern (2014) argue that claiming the Eugene Goostman chatbot passed the Turing Test is fundamentally absurd as Turing's prediction that in 50 years conversational AI could pass as a human was merely hypothetical, akin to a statistical survey or Gallup poll. Turing's acumen is a methodology to explain how the human mind functions by developing a computer capable of proximal behaviour and intelligence, which includes verbal processing and sensorimotor/robotic dimensions in which AI is systematically grounded (Sample & Hern, 2014).

In consideration, Harnad (2000) argues that the Turing Test is not a measure of how an AI system operates over five minutes; it is the system's ability to simulate the human mind over a lifetime. According to Gehl (2013), a similar text-based chatbot named Cleverbot claimed to pass the Turing Test in 2011 at the Technie festival in India, four years before the Eugene Goostman chatbot. However, Cleverbot did not receive the media coverage and scholarly attention of the Eugene Goostman program due to numerous irregularities in the results.

Aron (2011), Jacquet et al. (2019) and Mann (2014) argue that although Cleverbot claimed to exceed the 30% benchmark of the Turing Test scoring an exceptional 59.3%, human interrogators rated human agents as AI at an even higher rate of 63.3%. Thus, significant discrepancies in the results indicate fundamental flaws in the evaluation procedure and recruitment process.

However, Landgrebe & Smith (2019), Jacquet et al. (2019) and Pereira (2019) argue that although numerous chatbot systems claim to pass the Turing Test. The modernised tests are weak variations of Turing's original proposition, which are not representative of Turing's hypothesis and therefore do not qualify as certified passes. Fawaz (2019) and Wakefield (2019) explain that creating chatbots to pass the Turing Test is a developer's past-time as there is no serious scientific research in developing AI to pass the Turing Test. In support, Sharkey (2012) suggest that as Turing is long deceased, clarifying the terms and conditions of passing the Turing Test is impossible.

In RHR design, Mori's (1970) UV accounts for the negative psychological stimulus propagated by RHRs upon observation, as the more human-like artificial humans appear, the greater the potential for humans to feel repulsed by their appearance. However, per Burleigh (2013), there are considerable arguments against the scientific value of the UV theorem, as many scholars regard it as purely academic. Thus, the UV like the Turing Test remains a controversial topic in AI and robotics.

2. The Turing Test

Alan Turing (1950) formulated the Turing Test to determine if a machine agent could mislead human interrogators into believing answers provided by a computer are those of a human. If the machine convinces 30%+ of human interrogators into thinking it is sentient, the system passes the test and the higher this percentage, the more humanistic the AI functions. Turing argues that if a machine agent is capable of exhibiting human behaviour indistinguishable to that of a human, then the artificial mind functions in a manner akin to the human mind (it can think). However, Turing questions a machine's ability to think as 'thinking' is problematic to define and thus proposes the Turing Test as a methodology to explore this concept.

Turing supposes that if a machine agent replaced either of the male or female agents in the imitation game and could operate with a level of intelligence proximal to the responses of a human, then it would replace his original hypothesis 'can machines think?' (Turing, 1950).In the Turing Test, the objective of the human interrogator is to identify which agent is AI and human by posing a series of questions to evaluate the authenticity of the responses to differentiate between the AI agent and human agent. It is the agent's role to deceive the human interrogator into believing that they are the opposite agent by providing type-written answers that simulate the responses of the other.

However, Turing applies constraints to the Turing Test to establish equilibrium between the agents. Firstly, Turing narrows the scope of interaction between the human interrogator and the human/machine agents to a single topic of conversation, to prevent the human interrogator asking questions outside of the scope of the AI system's capabilities which may allude to the artificiality of the system. Similarly, Turing restricts the human interrogator's ability to propose mathematical inquiries to the agents as machine's are capable of correctly answering complex equations consistently, unlike humans.

Secondly, Turing imposes a 15-30 second time delay between the responses of the human interrogator as machine agents require time to formulate and respond to questions, unlike the human mind to which responses are immediate. Thirdly, Turing limits the time-scale of the evaluation to 5 minutes to prevent the machine agent producing incorrect or repetitive responses as the longer the interaction, the higher the potentiality for error. However, Turing considers the physical emulation of the human being as a distraction from the pursuit of intelligent machine's (Turing, 1950, p.2). Although Turing is correct in stating that the appearance of a machine is not indicative of its intellectual capabilities, he neglects the capacity of the human body in tactile learning, socialisation and non-verbal communication which are vital processes in social learning and communication.

2.1 Arguments and Limitations of the Turing Test

In Searl's experiment, a human agent sat in the middle of a room is passed a series of random Chinese symbols from under a door. The agent uses an instruction manual to arrange the symbols to form coherent sentences. After a while, the agent becomes efficient in arranging the symbols into sentences and no longer requires the instruction manual. The instruction manual is removed and interrogators who are fluent in written Chinese observe the agent arrange the symbols into sentences and state whether they think the agent is literate in Chinese or not.

In the experiment, the interrogators agree that the agent is fluent in Chinese to form coherent sentences using the symbols. However, the agent only understands the order of the symbols and not their meaning and therefore, lacks the vital process of comprehension. Thus, the perception of the interrogators in Searl's experiment is critical in understanding how humans interpret the appearance of intelligent behaviour as in real-life conditions; there are no visual distinctions between functional intelligence and comprehension, visualised in Fig.1

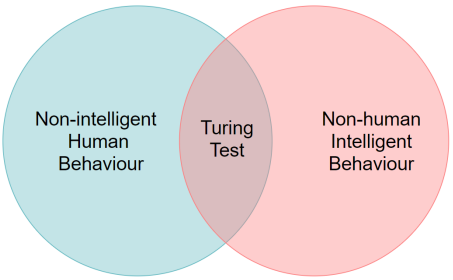

Fig. 1*: Visual representation of Searls Argument*

Similarly, Cole (2019) and Warwick & Shah (2015) argue that the Turing Test is susceptible to human interference by a fundamental design flaw which inverts the human perception of the nature of computing by remaining silent. Ghose (2016) explains that if an AI system does not answer questions when prompted, the human interrogators cannot distinguish between the silence of the AI and human responses; hence, the AI agent would pass as human by default. Thus, it is the expectancy for a computer system to respond to the actions of a human operator. If a computer does not perform tasks in a manner accustomed in HCI, this processual irregularity has the potentiality to influence human perception of the nature of the agent (Reynolds, 2016). In consideration, Hern (2019) and Landgrebe & Smith (2019), suggest that silence during the Turing Test is not uncommon and typically the result of poor programming.

However, stricter policies regarding the time limit of agent responses are crucial in maintaining the integrity of the Turing Test to irradicate purposeful exploitation of this loophole. Whitby (1996), argues that AI developer's and scholars have long misinterpreted the purpose of the Turing Test as Alan Turing designed the 'Imitation Game' as a game and not a formal test. Whitby argues that Alan Turing never intended the imitation game as an evaluation of machine intelligence, but rather as a thought experiment for assessing a machine's capacity to portray the behaviours of a human authentically. Whitby suggests that Turing's paper is not an operational guide for AI, but a theoretical treatise to examine the sociological and scientific value of creating machine's which can mislead human beings into believing they are human. However, Whitby explains that simulating human personalities and emotion in AI is damaging as these attributes tend to be misleading rather than progress the intellectual capacity of AI. Thus, the practical value of Turing's hypothesis is not in creating machine's with intelligence proximal to humans known as artificial general intelligence (AGI), but in emulating the conditions of the human mind and behaviours using computers.

This concept is significant in HRI and HCI as it considers how humans interface and interact with technologies that simulate human intelligence, personalities and behaviour. Rapaport (2000) argues that the Turing Test is limited in its scope of evaluation as it only considers HCI via NLP. Stock-Homburg et al. (2020) describe the Handshake Turing Test (HTT) and similarly, Karniel et al. (2010) the Turing Handshake Test (THT) as tests to determine if human interrogators can identify the differences between a human and RHR by the act of a handshake (tactile HRI). Moreover, this approach neglects the emulation of appearance, communication, AI and movement by focusing on secondary aspects such as touch and temperature. Ishiguro (2005) developed the Total Turing Test (TTT) for RHRs in HRI, formulated on Harnad's (1992) TTT for human-computer interaction and Harnad's (2000) Robot Turing Test (RTT) to comprehensively evaluate the appearance, behaviour and movement of RHRs against a human counterpart. Ishiguro's (2005) TTT implements point of view (POV) cameras mounted on the heads of the human and RHR agents. The agents conduct logistical tasks, and it is the role of the human interrogator to discern which agent is human and RHR from observation. Secondly, the human interrogator observes live 'full body' video streams of the agents for two seconds and decides which agent is human and RHR. Kasaki et al. (2016) cite 70% of subjects identified the movements of RHRs as human. Ishiguro argues that the Turing Test evaluates the intellectual capabilities of a computer on the assumption that the human mind is divisible from the body.

Thus, the TTT evaluates embodied artificial intelligence (EAI) by combining intelligent behaviour with a robotic body for assessing the human likeness of robotic behaviour, appearance and movement. However, the TTT is susceptible to design flaws; Firstly, live video footage is inaccessible. Secondly, Marzano & Novembre (2017) argue that the 2-second evaluation window is too limited. Thirdly, according to Schweizer (1998) & Bringsjord et al. (2000), the TTT is not a comprehensive approach as it neglects the evaluation of NLP to robotic mouth articulation during HRI. Fourthly, Oppy (2003) stipulates that judging the authenticity of intelligent behaviour by manipulating objects is not indicative of a machine's intellectual capacity. In consideration, Schweizer (1998) created the Truly Total Turing Test (TTTT) to remove telepresence from the TTT and evaluate automated RHR's with EAI. However, the TTTT lacks vital processes such as physical examination, movement, appearance, materiality, EAI and communication when operating as one robotic system.

3. The Multimodal Turing Test

Per the findings of the literature review, current evaluation methods used to determine degrees of human likeness in RHRs in HRI and HCI, such as The Turing Test, TTT, TTTT, RTT, THT and HTT are too limited in their scope of evaluation as they neglect the significance of amalgamating; communication (speech and gesturing), movement, vision, aesthetics and conversational AI into a single system, which is not representative of the human condition. In consideration, this study lays the foundations of a comprehensive theoretical evaluation methodology named the Multimodal Turing Test (MTT) to determine if RHRs can attain a level of emulation perceptually indivisible from a human being, (Houser, 2019). As cited in a recent article in the Guardian UK, the MTT is more holistic than the original Turing Test, and previous evaluation methods in HRI by evaluating an RHRs appearance, communication, movement and AI (Mathieson, 2019), shown in Fig. 2.

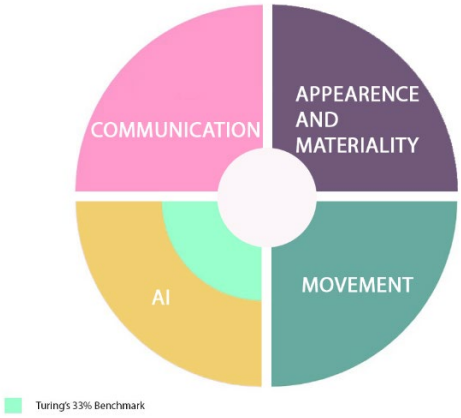

Fig. 2: The four evaluation modes of the MTT

The MTT incorporates the examination structure of the 1950 Turing Test by employing human interrogators to evaluate the perceptual authenticity of RHRs. However, unlike the binary pass / fail system of the original Turing Test, the MTT provides engineers, designers and programmers with a developmental framework to benchmark progress up to and in advance of Turing's 30% pass rate (Strathearn, 2019). Each stage of the MTT increases in complexity, which forms the hierarchy of human emulation shown in Fig. 3. Like Turing, it is not argued that an RHR metamorphosis into an organic system by replicating the conditions of a human being. However, if an RHR can appear and function in a manner indistinguishable from a human being in real-world conditions, then that RHR is perceptually indivisible from a living human being, The World Economic Forum (2019). Thus, equal consideration to the appearance and functionality of RHRs is essential to develop higher modes of human emulation.

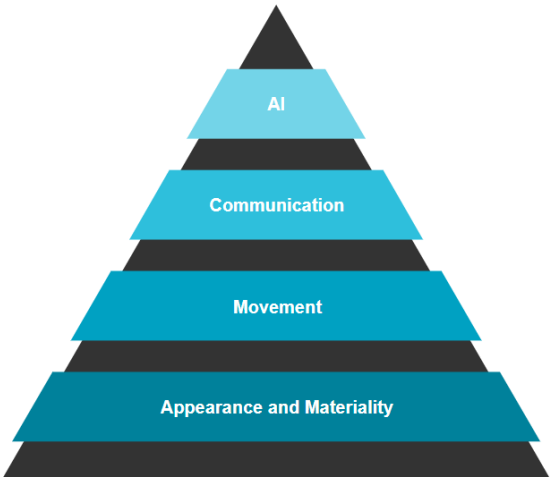

Fig. 3: The Hierarchy of MMT: Level 1 (Appearance), Level 2. (Appearance & Movement), Level 3 (Appearance, Movement & Communication) Level 4. (Appearance, Movement, Communication & AI).

However, replicating the appearance and materiality of a human is more straightforward than simulating human movement due to the complexity of natural kinetic variance. Therefore, per Baudrillard's (1994) order of simulacra, appearance forms the bedrock of the hierarchy of human emulation because it is the elementary form of simulation. Aesthetical appearance envelops a body to which movement is applied, as natural movement is more complicated to replicate than a still model; kinetics forms the second level of emulation. For speech to be a useful communication tool in RHRs, requires both an authentic appearance and naturalistic. AI is the apex of human emulation as the human mind is the most challenging element to simulate authentically due to its complexity. However, for AI to be a useful tool in RHRs, the emulated mind requires a human-like body and a method of communication for naturalistic HRI.

The four evaluation categories of the hierarchy of human emulation formulate a unified whole, which constitutes an RHR that can emulate (to degrees of likeness) a living human being, as reviewed in an article by Khatib (2019) which outlines the scope of the MTT. Furthermore, the MTT is an approach towards humanising forms of AI as current robotic AI predominantly focuses on logical, linguistical and kinesthetic intelligence and neglects interpersonal and intrapersonal intelligence to create higher modes of EAI. Interpersonal and intrapersonal AI is synergetic, incorporating various visual and audible stimuli such as facial expressions, vocal tonality, gesturing, and emotive responsivity to humanise AI interaction. This approach enhances the capacity for natural communication and responsivity between humans, and RHRs founded on authentically assimilating natural human-human interaction, (Barnfield, 2020).

Previous evaluation methods fall into the MTTs categories of human emulation, but none are inclusive of all four stages of development. For example, The THT and HTT, in movement (handgrip), the TTT falls under appearance, AI: Wizard of Oz (WOZ) method and kinetics (robotic vision, aesthetics and movement), the Turing Test in AI in (AI) and the TTTT in appearance, movement and AI. However, developing an RHR as a complete system with components across all four categories of the hierarchy of human emulation (without consideration of the stages) will not achieve levels of human likeness indivisible from a human being. For example, comparing two RHR heads to determine which one is more visually authentic than the other is a viable methodology for evaluating and testing new components by increasing the realism of one robotic head over the other.

However, this approach is futile when comparing RHRs against a living human being to determine authenticity as the distinctions in form and function are highly apparent, as exemplified in Fig. 4.

Fig. 4: RHRs developed in this study / Human Comparison. Left: RHR, Baudi. Middle: RHR, Euclid. Right: Human head

Therefore, a multimodal approach is required using a controlled evaluation methodology by combining features that belong to the same body (subgroup), such as, EAI, natural speech synthesis and a robotic jaw, tongue and lips. This evaluation procedure applies to other subgroups such as eyes: (sclera, pupil dilation, iris, eyelid, eyelashes, veins, eye movement, blink rate, skin, hair, aesthetics) and so on. This approach is similar to the functional constraints of the original Turing Test to control the direction and flow of a conversation by narrowing it to a specified theme or topic of discussion. This technique permits the refinement of smaller intricate motor functions and aesthetics within the subgroups, indicated in Fig. 5.

Fig. 5: *Mouth comparison* Left: RHR. Right: Human

The MTT is a method for overcoming many of the design issues that are prevalent in RHRs such as inaccurate eye emulation, poor aesthetical design and unnatural movement. Furthermore, according to the uncanny valley hypothesis, realistic humanoids instigate negative perceptual feedback in humans because they are void of variable organic nuances.

This consideration is vital in the development and progression of modern RHRs, as traditional methods of evaluation and design overlook the significance of replicating nuances such as pupil dilation, gestures and accurate lip movement. These facial expressions act as visual cues and signifiers of sentience when discerning the authenticity of an RHR.

Thus, when evaluating an RHR, all elements are interconnected to the perceptual whole. To achieve this, an imitation head structure and cloaking device to cover empty areas around the developed feature is a practical method of resolving this issue. This approach permits a holistic evaluation compared to analysing individual facial features outside of the body (unified whole).

The Multimodal Turing Test: three orders of human emulation: The three orders of human emulation are a framework for developing RHRs that appear and function in a manner that is indistinguishable from the natural human being under the conditions and limitations of the MTT evaluation procedure.

1. Fragmentary Emulation: A unified subgroup that qualifies as perceptually indistinguishable in form / and or function when compared to a human.

2. Synchronised Emulation: A set of two or more subgroups that are perceptually indivisible in form / and or function from a living human being.

3. Absolute Emulation: A fully assembled human replicant consisting of all subgroups working as a unified whole to emulate the human form and function.

The total length of the MTT is 20 minutes and divided into four 5-minute evaluation sections, covering: appearance, movement, voice and AI founded on the five-minute evaluation rule of the original Turing test. The MTT has broader applications outside the field of RHRs and EAI in realistic virtual humanoids (RVHs) with EAI for HCI. Developing higher modes of human likeness in RVHs is significant in EAI interface design for HCI and exploring the UV in RVHs. Therefore, it is essential to provide evaluation conditions for assessing the perceptual authenticity of RVHs for the future progression of virtual humanoids towards a simulacrum indivisible from living humans.

4. The Multimodal Turing Test for RHRs

The MTT is more comprehensive than the Turing Test, TTT, RTT, THT and HTT by systematically examining appearance, functionality, AI and voice processing to provide a universal evaluation procedure for all types of humanoid robots with varying degrees of human likeness. This multimodality requires several constraints to ensure the integrity of the evaluation procedure. In Fig. 6, the human Interrogator (A) evaluates the authenticity of agents (B) and (C) who are separated by a solid screen to minimise interference. Significantly, both agents (B) and (C) inhabit the same physical environment and visual spectrum as the human interrogator for greater perceptual authenticity.

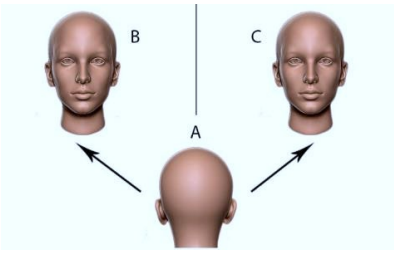

Fig. 6: MTT for RHRs Evaluation Environment. A: Human Interrogator. B: Human / RHR Agent C: Human / RHR Agent

4.1 First Stage: Appearance

The first stage of the MTT requires human interrogator (A) to evaluate the appearance of agents (B) and (C). Different subgroups contain different visual elements such as lips, hair, skin tone and wrinkles. Therefore, imperfections in synthetic skin such as wrinkles, spots and blemishes are essential as these defects are not typically associated with RHRs. The first level examines the visual authenticity of the agents, such as an area of natural skin of Agent (B) with the corresponding synthetic skin area from Agent (C). The MTT is significant to the progression of RHRs as the Turing Test does not provide a developmental framework due to the binary pass/fail system.

Thus, allowing engineers to gauge the authenticity of specific facial/bodily areas individually, as a group, or as a complete form towards attaining the pass threshold (emulation that is indivisible from a living human) is essential. It is crucial to evaluate Agent (B) against (C) and then Agent (C) against (B) for a detailed and comprehensive analysis. For example, imagine Agent (B) is a robotic mouth and (C) a human mouth, and the human interrogator (A) identifies a visual irregularity in the bottom lip of Agent (B) leading to the human interrogator identifying Agent (B) as an RHR. This process applies to every item within a subgroup to pinpoint the precise location of the visual irregularity. It is vital to access the aesthetical quality of the inside of the robotic mouth during the first stage evaluation as this area is exposed during operation.

4.2. Second Stage: Movement and Dexterity

The second stage of the MTT incorporates both movement and appearance; The human interrogator (A) selects an expression or gesture from a list of commands, such as smile, frown, wave, open mouth. The Human interrogator (A) selects which agent performs the command by addressing the agent and saying aloud the command. As in the Turing Test, a delay in the response time (5-10s) of the agents allows time for NLP. Servomotor sounds must be triggered by the human agent when performing physical movements to reduce signifiers such as sound interference that may allude to the mechanical nature of the RHR. It is essential to assess tongue movement to match vowel and consonant sound as the internal components are exposed by the robotic mouth during verbal communication. The accurate replication of acute motor functions such as pupil dilation, breathing, facial tics and blink rate must be considered in the second stage. Furthermore, the complexity and level of movement are variable on the style of the humanoid robot; for instance, robotic heads do not require the evaluation of body movement such as hand gesturing.

However, evaluating hand gestures is essential for a ‘waist up’ robot design. Comparatively, a waist up robot does not require the evaluation of leg movement and balance, unlike a full-body humanoid robot which needs the robot to stand and move the lower parts of its body. Therefore, applying constraints to control the evaluation area for different styles of RHRs is significant, for example; seating robotic heads and waist-up robots and at a table during the evaluation procedure will reduce and concentrate the evaluation area. This method is standard in HRI to conceal an RHRs lower body and external mechanical components from the observer. If an RHR can pass the first two stages of the MTT at a rate of 30%+, is the same as saying in real-world conditions, an RHR is visually indistinguishable from a living human being (without speaking or AI interaction).

4.3 Third Stage: Speech and Mouth Articulation

The third stage of the MTT evaluates an RHRs speech, lip dexterity and aesthetical appearance. It is not the objective of the MTT to develop a more human-sounding robotic voice as this field is continually evolving outside of RHR design. However, the MTT examines the compatibility and accuracy of speech synthesis with robotic mouth articulation. Speech synthesis technologies are advancing rapidly and continually improving in human likeness, and the use of current and future speech synthesis technologies in RHRs is significant towards total automation.

Using NLP in the MTT is preferable to human speech as it protects the integrity of the test environment by seamlessly interchanging between the previous evaluation stages. However, as speech synthesis is yet to replicate human speech, implementing current speech synthesis is counterproductive when developing RHRs that are perceptually indivisible from humans. Therefore, it is essential to outline an alternative methodology of natural speech processing to overcome the current limitations of computerised speech technologies. The WOZ approach permits a second human agent (D) to speak in place of the robotic voice, as demonstrated in Fig. 7. The speech of Agents (D) and (C) are relayed to the human interrogator (A) by headphones to minimise the sound difference between the speaker system and natural human voice.

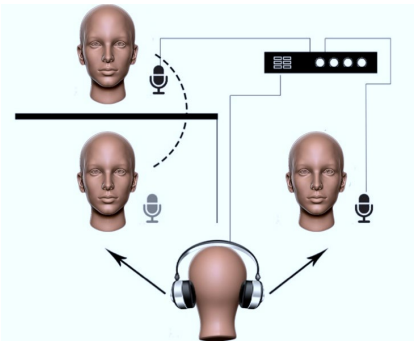

Fig. 7: Natural voice to a robot speech/mouth actuation

This approach permits the examination of human speech using a robotic mouth system, allowing for a greater accurate comparative evaluation than current speech synthesis. However, real-time human speech to lip synchronisation is less reliable than speech synthesis due to the variability in pitch, volume, frequency and tonality of human speech. Therefore, it is essential to configure the robotic mouth to function with one human voice for optimum lip-synchronisation accuracy. Although the evaluation for natural human speech and computerised speech is different, the procedure is identical. The human interrogator (A) engages in an interactive game with agents (B+D) and (C). The objective of the game is for the human interrogator (A) to guess what animals that agents (B+D) and (C) are thinking of by posing questions to each of them about the animal′s appearance, habitat, movement and diet. The human interrogator (A) rates and compares the authenticity of Agents (B) and (C) voice and mouth articulation. This approach is vital for evaluating speech, as implementing a structured gamification methodology does not require deep learning or machine learning methods and permits the human interrogator to focus on speech quality rather than correct or incorrect AI responses. Finally, time limitations on ‘silence’ are significant to upholding the integrity of the MTT as suggested in an article on the MTT and the Turing Test, (Cole, 2019).

Therefore, a time limitation of 10 seconds is imposed and strictly monitored throughout the evaluation procedure, with time added to the end of each session if silence is excessive or exceeds the 10-second maxima. If an RHR can pass the third stage of the MTT, then that systems autonomous speech processing and tonal expressions are proximal to natural human speech and mouth/lip movement, facial expressions and appearance. However, for an RHR to progress to the final stage (AI) of the hierarchy of human emulation, the system must be fully automated without human control for the integration of speech and AI. Therefore, implementing the alternate speech evaluation procedure is an acceptable method for passing the third level of the MTT but not for progressing onto the final level.

4.4 Final Stage: AI (Absolute Emulation)

The final stage of the MTT is inclusive of all four elements: intelligence, movement, speech and appearance. It is vital at this stage that all human control is removed, permitting the RHR to function autonomously and the AI to control the operations of movement and speech. As EAI constitutes the ‘personality’ of the RHR, developer’s need to create an AI people personality with interests and traits that match the appearance, speech synthesis and movement of the RHR. Passing the final stage of the MTT would answer the question: can machines emulate a human being? Therefore, developing an EAI program to control accurately trigger facial expressions, voice tone, emotions and gestures are crucial in the final evaluation. This method is the foundation for developing more sophisticated modes of interpersonal AI for robots. Like the Turing Test, the final stage evaluation focuses on a single topic of discussion selected by the human interrogator from a pre-established list of subjects. The final test lasts 5 minutes with the human interrogator (A) posing 2.5 minutes of questioning to agents (B) and (C) on the selected topic. As technology improves NLP, RHR and AI efficiency, this time limit should be extended until the RHR can deceive a human interrogator indefinitely. At the end of the evaluation procedure, the human interrogator (A) chooses which agent (B) or (C) is human (or unsure) and provide a detailed account of the decision-making process covering all evaluation categories. If 30%+ of test subjects misidentify or are unable to discern the difference between the RHR and the human agent, then the RHR has succeeded in passing the final stage of the MTT However, if an RHR does not pass all stages of the MTT, the data gathered during the test stages will provide engineers with information concerning specific area/s that emit irregular feedback through the layered evaluation process for revision or calibration.

5. Conclusion

The MTT is an essential evaluation method towards achieving higher modes of human likeness in RHRs and EAI as in other methods of evaluation; slight miscalculations of an otherwise realistic-looking robot can allude to the robot's artificiality resulting in other high-quality components becoming part of that failure. The objective of the MTT is to permit engineers to work systematically and build up areas of the face and body to ensure all components are equal to that of a human before expanding the fields and adding more features towards creating a complete RHR that is perceptually indivisible from a living human being.

References

AAmoth. D (2014) The Fake Kid Who Passed the Turing Test. Ret:time.com/2847900/eugene-goostman-turing-test/.Acc 5.2.20

Aron. J (2011) AI tricks people into thinking it is human. Retrieved: newscientist.com/article/dn20865-software-tricks-people-into-thinking-it-ishuman/#ixzz6Ez0JgQi1. Acc: 25.02.20

Barnfield. N (2020) Face to Face With The Future of AI. *Horizon Magazine*. Riley Raven. DOI: https://www.staffs.ac.uk/alumni/ horizon-alumni-magazine pp.8-11

Baudrillard. J (1994). Simulacra and simulation. Trans: Ann Arbor : University of Michigan Press, ISBN-10: 0472065211.

Bringsjord, S., Caporale, C., & Noel, R. (2000). The Total Turing Test, *J-LLI, 9*(4), 397-418. DOI:www.jstor.org/stable/40180234

Burleigh. T, Schoenherr. J, Lacroix. G (2013). Does the uncanny valley exist? Computers in Human Behaviour. 29. 759-771. DOI:10.1016/j.chb.2012.11.021.

Cole. E (2019) What is New in Robotics? Retrieved: blog.robotiq.com/whats-new-in-robotics-06.12.2019. A:25.02.20

Copeland. J (2014) Why Eugene Goostman Did Not Pass the Turing Test. Retrieved: https://www.huffingtonpost.co.uk/jack-copeland/turingtesteugenegoostman. Acc:25.02.20

Fawaz. A (2019) A tangible Turing Test. Retrieved: https://www.neowin.net/news/a-tangible-turing-testthe-loebner-prizeis-coming-to-swansea-this-weekend/. Acc: 21.02.20

Gehl. R (2014). Teaching to the Turing Test with Cleverbot. Transformations: The Journal of Inclusive Scholarship and Pedagogy, 24(1-2), 56-66. Retrieved February 25, 2020.

Ghose. T (2016) Robots Could Hack Turing Test by Keeping Silent. Retrieved: www.scientificamerican.com/article/robots-could-hack-turing-test-by-keeping-silent/. Acc: 25.02.20

Harnad, S. (1992) The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion. SIGART Bulletin 3(4) (October 1992) pp. 9 - 10.

Harnad, S. (2000) Minds Machine's and Turing, *Journal of Logic, Language and Information*, vol 9: p.425. https://doi.or g/10.1023/A:100 8315308862

Hern. A (2014) What is the Turing Test? Retrieved: www.theguardian.com/technology/2014/jun/09/what-is-the-alan-turing-test. Acc: 25.02.20

Houser. K (2019) Advanced Robotics Forced Scientist To Invent A New Turing Test. Retrieved: https://futurism.com/the-byte/scientists-invented-new-turing-test. Acc: 18.04.20

Ishiguro, H. (2005). Android science: Toward a new cross-interdisciplinary framework. J-Comp-Sci, Corpus ID: 6105971Jacquet, B., Baratgin, J., & Jamet, F. (2019). Cooperation in Online Conversations. Journal of Psychology, 10, 727. https://doi.org/10.3389/fpsyg.2019.00727

Karniel, A & Avraham, G, Peles, Ba & Levy-Tzedek (2010). Turing-Like Handshake Test for Motor Intelligence. *JoVE.* 10.3791/2492.

Kasaki, M. & Ishiguro, H. & Asada, M. & Osaka, M. & Fujikado, T. (2016). Cognitive neuroscience Robotics: Synthetic Approaches to human understanding. 10.1007/978-4-431-54595.

Khatib. H (2019) Just because they are robots? Retrieved: www.ameinfo.com/industry/technology/robots-treat-humanoids-racial-gender-bias Acc: 15.09.19

Landgrebe. J Smith. B (2019) There is no AGI. Retrieved: https://arxiv.org/abs/1906.05833. Acc: 25.02.20

Mann. A (2014) The computer actually got an F on the Turing Test. Ret: wired.com/2014/06/turing-test-not-so-fast/. Acc:9.2.20

Marzano, G & Novembre, A. (2016). Machine's that Dream: A New Challenge in Behavioral-Basic Robotics. Procedia Computer Science. 104. 146-151. 10.1016/j.procs.2017.01.089.

Mathieson. S (2019) Will androids ever be able to convince people they are human? Retrieved: www.researchgate.net /publication/341756234_Mr_Robot_Will_androids_ever_be_able_to_convince_people_they_are_human_Guardian_Online. Acc 07.07.20

Mori. M (1970). The Uncanny Valley. Energy, Issue 7, pp.33-35. DOI: 10.1109/MRA.2012.2192811

Oppy, G. R., & Dowe, D. L. (2003). The Turing Test. *Stanford Encyclopedia of Philosophy*, 1(online), 1 - 26.

Pereira. D (2019) You should fear Super Stupidity, not ASI Retrieved:towardsdatascience.com/you-should-fear-super-stupidity-not-super-intelligence-19f93a46fa4d. Acc.17.02.20

Rapaport, W. (2000). How to Pass a Turing Test. *Journal of Logic, Language, and Information*, 9(4), 467-490. Retrieved: www.jstor.org/stable/40180238. Acc: 24.04 2020

Reynolds. E (2016) Does the Fifth Amendment 'expose a serious flaw' in Turing Test? Retrieved:www.wired.co.uk/article/major-flaw-turing-test-silence. Acc:25.0 2.20

Robbins. M (2014) A Machine Did not 'Pass' the Turing Test. Retrieved: https://www.vice.com/en_uk/article/gq8ddw/eugene-goostman-alan-turing-test-kevin-warwick. Acc: 25.02.20

Sample. I & Hern. A (2014) Scientists dispute if 'Eugene Goostman' passed Turing Test. Retrieved: www.theguardian.com/technology/2014/jun/09/scientistsdisagree-over-whether-Turing-test-has-been-passed. Acc. 19.04.20

Schweizer. P (1998). The Truly Total Turing Test. Minds Mach. 8, 2 (May 1998), 263–272. DOI:10.1023/A:1008229619541

Searle, J (1980). Minds, brains, and programs. Behavioural and Brain Sciences 3 (3): 417-457, DOI: 10.1.1.83.5248.

Sharkey. N (2012) Alan Turing: The experiment that shaped AI. Retrieved: bbc.co.uk/news/technology-18475646. Acc:22.02.20

Stock-Homburg. R, Peters. J, Schneider. K, Prasad. V, Nukovic. L (2020) Evaluation of the HTT for anthropomorphic Robots, *Int Conf HRI.* DOI: 10.1145/3371382.3378260

The World Economic Forum (2019) Can machine's think? A new Turing Test may have the answer. Retrieved: www.weforum.org/agenda/2019/08/our-turing-test-for-androids-will-judge-how-lifelike-humanoid-robots-can-be/. Acc: 14.03.20

Turing, A. (1950). Computing Machinery and Intelligence, *Mind,* (236), pp.433-460. doi.org/10.1093/mind/LIX.236.433.

Wakefield. J (2019) The hobbyists competing to make AI human. Retrieved:www.bbc.co.uk/news/technology-49578503 Acc: 25.02.20

Warwick, K., & Shah, H. (2015). Passing the Turing Test Does Not Mean the End of Humanity. Cognitive Computation, 8, 409-419. DOI: 10.1007/s12559-015-9372-6

Whitby, B (1996) Reflections on AI: the legal, moral and ethical dimensions. Intellect, Oxford, UK. ISBN 9781871516685