Noname manuscript No.
(will be inserted by the editor)

A Novel Speech to Mouth Articulation System for Realistic Humanoid Robots

Carl Strathearn · Minhua Ma.

the date of receipt and acceptance should be inserted later

Abstract A significant ongoing issue in realistic humanoid robotics (RHRs) is inaccurate speech to mouth synchronisation. Even the most advanced robotic systems cannot authentically emulate the natural movements of the human jaw, lips and tongue during verbal communication. These visual and functional irregularities have the potential to propagate the Uncanny Valley Effect (UVE) and reduce speech understanding in human-robot interaction (HRI). This paper outlines the development and testing of a novel Computer Aided Design (CAD) robotic mouth prototype with buccinator actuators for emulating the fluidic movements of the human mouth. The robotic mouth system incorporates a custom Machine Learning (ML) application that measures the acoustic qualities of speech synthesis (SS) and translates this data into servomotor triangulation for triggering jaw, lip and tongue positions. The objective of this study is to improve current robotic mouth design and provide engineers with a framework for increasing the authenticity, accuracy and communication capabilities of RHRs for HRI. The primary contributions of this study are the engineering of a robotic mouth prototype and the programming of a speech processing application that achieved a 79.4% syllable accuracy, 86.7% lip synchronisation accuracy and 0.1s speech to mouth articulation differential.

Keywords Realistic Humanoid Robotics · Lip Synchronisation · Human Robot Interaction · Machine Learning

F. Author
Edinburgh Napier University
E-mail: c.strathearn@napier.ac.uk

S. Author
Falmouth University
E-mail: m.ma@falmouth.ac.uk

1 Introduction

Many scholars consider the creation of an RHR that is perceptually indistinguishable in appearance and functionality to that of the average human as the apex of mankind's technological achievements [1] & [2]. However, no RHR is capable of convincingly emulating the human condition due to the complexity of the problem [3]. A key failure in RHR design is accurately synthesising the appearance, speech, movement and intelligence of RHRs to function naturally in the real-world [4]. This consideration is significant as the longer the interaction between humans and RHRs, the greater the probability for visual and functional irregularities to materialise and allude the robot's artificiality [5]. Thus, as the mouth area is the primary focal point of communication in face to face interaction, inaccurate speech synchronisation, vocal tonality, mouth aesthetics and movement significantly reduces natural speech reading [6] and speech understanding [7] in HRI. Excessive time differentials between speech and mouth articulation occurs primarily in NLP systems that rely on vowel and consonant pattern extraction from text to control lip movement and SS due to the demanding system loads [8]. Comparatively, audio signal dependent lip synchronisation applications have a higher response time, but they are not as precise as text processing methods due to the highly variable sound waves in natural voice output [9]. Thus, response time and accuracy are crucial factors in measuring the authenticity of real-time lip synchronisation systems. An advantage of audio-signal processing over text extraction is the ability to implement human speech in place of a SS, which is a common practise in contemporary RHR design.

For instance, the Wizard of Oz (WOZ) approach is typically implemented in RHRs as SS applications are incapable of accurately emulating the natural vocal tones of human speech [10]. However, creating RHRs that function autonomously in real-world environments is more significant in HRI as they can perform tasks in highly variable conditions without the need of a human operator. Furthermore, natural tongue positions during speech are often overlooked in humanoid robot design, which reduces the natural appearance of the system rather than its speech functionality as the robotic tongue is not required to formulate word sounds like a human tongue. However, aesthetical accuracy is crucial to maintaining the perceptual authenticity of the humanoid robot by reducing the UVE.

2 Human Mouth Muscle Configuration

The human mouth consists of a fine network of muscular fibres surrounding the upper and lower jaw structure of the mouth and cheek/neck bones [11]. The, orbicularis oris muscle group indicated in Fig .1, is responsible for the pursing and stretching of the lips for forming vowel and consonant sounds which is essential in the development of a realistic robotic mouth.

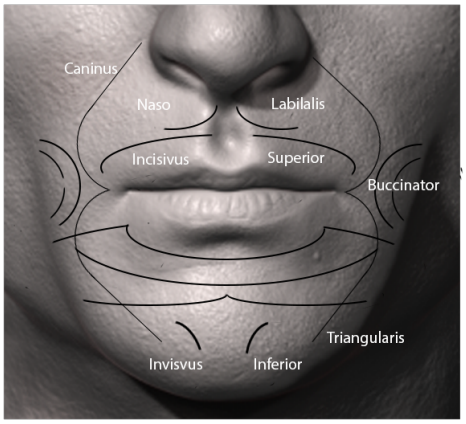

Figure. 1 The Orbicularis Oris Muscle Group

The thin quadrilateral muscle fibres help form lip shapes to enunciate vowel and consonant sounds. For example, O, U and Q are known as rounded word sounds which require the buccinator muscle to protrude forward to form, shape and phonetically pronounce words that start with O, U and Q. The buccinator muscle group is also significant in forming facial expressions, such as smiling and frowning. Thus, replicating this muscle group is vital in developing a robotic mouth that can accurately emulate rounded word sounds and human facial expressions. However, as indicated in Figure .1, the buccinator has a crispation motion as well as a horizontal stretching motion, which actuates the lips inwards and outwards. The caninus muscles lift the corners of the mouth, and the triangularis muscles lower them; these muscle groups are essential design considerations when building a robotic mouth as the space between the upper and lower jaw is limited.

Therefore, the internal mechanism controlling this function must be capable of performing multiple points of articulation within the confines of the area between the teeth and skin of the robot. If the mechanism operates outside the parameters of this area, its movement will be visually noticeable on the surface of the silicone skin and may lead to eventual wearing and tearing of the skin. The incisivus and orbicularis superior muscle groups control the upwards, and downwards motion of the top lip and the invisvus and orbicularis inferior muscles control the upwards and downwards action of the bottom lip. This group of muscles are responsible for the opening and closing of lips which work independently from the lower jaw mandible. The incisivus, orbicularis superior, orbicularis interior and invisvus muscles work in synthesis with the buccinator muscles to form a complex array of lip positions to pronounce words.

However, although the incisivus and orbicularis muscles can work independently of one another the invisvus and orbicularis inferior work conjointly to form a singular motion, therefore, emulating the bottom lip muscle functions of the human mouth requires only a single actuator to perform the motion of the invisvus and orbicularis muscles.

The naso and labilalis muscles groups are made from denser sinew towards the base of the nose that forms the philtrum. Although these muscles create an indentation between the mouth and nose known as the medial cleft, these muscles are not vital in the development of a robotic mouth as the servomotors that emulate the actuation of the incisivus and orbicularis superior muscles are capable of lifting the silicone skin of the robotic mouth without additional support. However, although the naso and labilalis muscles are not required to function in a robotic mouth, the replication of the muscles under the skin is aesthetically significant to creating a realistic humanoid robot. Therefore, the muscular protrusion of the medial cleft should be visible in the silicone skin on the mouth of the robot, even if it is non-functional. This method is vital in maintaining the perceptual realism of the robots as even slight deformations in the appearance and functionality of a humanoid robot has the potentiality to propagate the UVE.

3 The Uncanny Valley

M. Mori [12] formulated the Uncanny Valley (UV) hypothesis to explain the causal effects of negative perceptual stimulus propagated by RHRs. However, many scholars consider the UV a transitionary theorem that is viable up to the point when engineers have the technology and tools to precisely replicate the human condition [13]. As advocated in a recent study [14] "We believe that a fear of the uncanny valley is unwarranted, and even potentially detrimental to the pursuit of design goals where human likeness is involved". Furthermore, key texts that support the UV rely solely on anecdotal evidence rather than scientific data [15] and many HRI studies employ Realistic Virtual Humanoids (RVHs) rather than RHRs which is not representative of real-world conditions [16].

Recent research [17] measured gaze frequency when examining RVHs faces and determined that the dwell time for determining the authenticity of the eyes (30-65s) was higher than the mouth (10-15s). However, the experiment utilised still imagery of RHRs rather than practical moving systems. Therefore, as eyes are greater in aesthetical detail than the human mouth, the dwell response time in determining their authenticity is substantially higher, yet, this study does not account for the movement of facial features. Conversly, [18] Grimshaw et al. (2011 determined that the human mouth is the most significant facial feature for reading and emitting recognisable emotions and a lack of articulation in jaw and lip movement heightened the UVE. Similarly,

[19] Tromp et al. (1998),[20] Nass et al. (2000) and [21] Garau et al. (2003) collectively conclude that functional realism is as crucial as aesthetical realism in RHR design. Therefore, determining the primary facial feature which emits the UVE is seemingly dependant on the type of evaluation procedure (still image or moving video). In language understanding, McGurk [22] (1976) coined 'the McGurk effect', which is the influence of visual speechreading and the spoken word during speech-visual communication. For example, if a person pronounces the word 'ba' but the speaker mouths 'ga' then there is a high potentiality for that person to hear either 'ba' or 'ga' or neither. According to [23] Ciechanowski (2018), there is an UV for robotic voices as although speech synthesis applications produce high-quality human voice simulations; they lack authentic human tonality and intonation giving them a distinctly robotic quality. This study highlights a gap in current knowledge as little practical research into the significance of accurate mouth and speech design is available due to the broader use of RVHs when examining the UV.

4 Natural Language Processing and Acoustic Waveform Analysis

NLP consists of 4 elements: Automated Speech Recognition (ASR): speech-to-text, Natural Language Understanding (NLU): decoding input, Natural Language Generation (NLG): structuring coherent sentences and Speech Synthesis (SS): text-to-speech. ASR and SS are not part of the intellectual capacity of AI systems as they are not elements of machine comprehension or learning [24]. Therefore, ASR and SS applications are suitable for robotic systems as they do not require human control. Furthermore, the real-time data input from SS applications into microprocessors to splice and analyse the acoustic qualities of the incoming audio are more stable and accurate than real-time human voice via a microphone. The difference in sound/voice quality is due to the impact of environmental, gender, age and stress conditions of the human operator compared to computerised speech which is unchanged by these factors. For example, a recent study [7] describes a novel approach to analysing vowel sounds from live audio by interpreting the hertz (Hz) frequency of incoming speech from a microphone input using an energy-based vocal tract model to formulate lip position for virtual characters. However, the study details the limitations of the lip-synchronisation application when decoding live input as the system neglected to recognise the variability in the tonality and pitch of different human voices which frequently produced incorrect lip positions. A similar study [25] explored the implementation of speech wave signal processing to create jaw movement in robots, derived from previous studies by [26] and [27]. Although the mouth articulation system was successful in correctly analysing incoming audio to detect frequency on/off status for synchronisation with the open/closed mouth positions of the robots, this study and the previous examples neglect lip synchronisation and focus solely on jaw position to incoming sound frequencies.

Thus, although this methodology is vital in the development of a robotic mouth system as it accounts for jaw positioning to syllable pattering and pitch frequency it requires further modification to include lip articulation for generating vowel/consonant sounds. This mouth configuration is significant, as discussed previously, immobile or muted robotic lip actuation has a high potential to be interpreted as aggressive or unemotional and generate negative perceptual stimulus. In support, [28] developed a robotic mouth system to examine common lip-syncing factors affecting humanoid robots. However, the study claims that the robot developed for this research can perform complex mouth shapes for replicating human vowel and consonant lip patterns. Yet, on review of the visual data provided in this research, the robot appears to lack genuine mouth actuation of the buccinator muscle group. This configuration is restrictive of the pursing and stretching of the corners of the mouth, which is a necessity when forming vowel and consonant lip. Furthermore, the response time of the system appears to be highly variable with ranges between 0.4 and 3.14s, which is extensive compared to human-human communication. The results of this study highlight these issues, i.e. "factors affecting communication with an android robot were mouth shape and lip-synced timing".

Thus, consideration of the buccinator muscle in the development of an accurate robotic mouth is crucial, as neglecting the actuation of the corners of the mouth is insufficient for creating accurate lip positions. A comprehensive study by [29] Ailm and Rashind (2018) examined the potentiality of common methods of phoneme and acoustic features extraction for formulating accurate lip positions, including Mel Frequency Cepstral Coefficient (MFCC) for identifying monosyllabic words, Linear Prediction Coefficient (LPC) which approximates formants and reduces signal noise to estimate frequencies of the vocal tract, Linear Prediction Cepstral Coefficient (LPCC) analyses vocal waveform frequency patterns, Line Spectral Frequencies (LSF) examines audio input into two filters to determine if the vocal tract is open or closed to define lip shapes, Discrete Wavelet Transform (DWT) examines the time and frequency patterns in speech for high-frequency events identification and Perceptual Linear Prediction (PLP) analyses the pitch or loudness of speech frequencies known as the 'bark' approach. However, the study concludes that although these methods have stood the test of time, they are still susceptible to incorrect data generated by variable speech patterns in user input such as accents, age and gender. A recent speech to video application for virtual characters by [30] implements the MFCC method of audio signal analysis and derived compelling results. The application has a 0.35s response time to generate lip synchronisation patterns for the virtual characters, which is higher than [7] model of 0.5s and [28] 0.4s-3.14s robotic mouth response time . Thus, the MFCC approach is significant in developing a robotic mouth that can respond to incoming live audio transmissions due to its high speed of computation and a low field of noise interference.

5 The Design and Programming of the Novel Robotic Mouth System

The development of the robotic mouth created in this study combines virtual rendering, virtual simulation evaluation, 3D printing, CNC machining and traditional engineering techniques. The development process initiated with a virtual rendering created in Computer-Aided Designed (CAD) software Solidworks 2019 using the muscle position and movement chart created in the literature review, depicted in Figure.2.

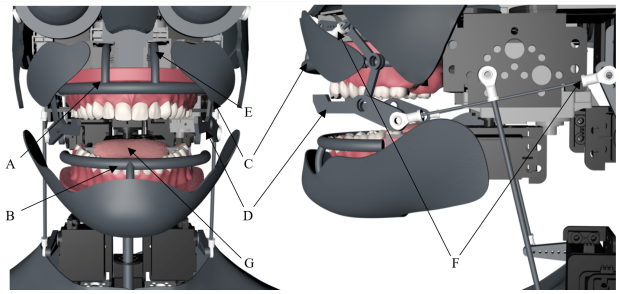

Figure. 2 CAD Schematic of robotic mouth. Front. A: Mechanism to replicate the incisivus muscle. B: Single mechanism to control the up/down motion of the bottom lip to emulate the invisvus and orbicularis inferior muscle groups. C: Left and Right cheek components. D: Actuator to emulate the buccinator muscles. E: Mechanism to replicate the orbicularis major muscle. F: Servomotor direct drive train to operate the left/right stretching of the mouth and pursing of the lips. G: Silicone Tongue with embedded actuators.

The parts list from Solidworks exports as .STL files into 3D animation modelling software Cinema 4D for texturing, colourisation and animation. A custom animation rig applied to the 3D model permits the evaluation and testing of the system when in motion. After successful testing and adjustments to the mechanisms, the parts export from Cinema 4D into Ultimaker Cura 3D printing software and printed on a CR-10s 3D printer using ABS plastic at a layer higher of 0.2mm and a temperature of 290°. However, during testing of the system, the mechanisms replicating the buccinator muscles fractured and splintered under pressure. Therefore, remaking this component in a stronger material became a necessity to handle the torque load. After several unsuccessful durability tests using different filaments such as PLA, SLA and 3D printable composite metals, non could withstand the forces of the servomotors due to the brittle layering process in 3D printing. Thus, a number of tests using non-3D printed metals proved that aluminium was the lightest and most durable material for creating the buccinator mechanism. However, as aluminium is not an extrudable material on a traditional 3D printer, this element had to be cut on a CNC lathe using the vector path exported from Solidworks, demonstrated in Figure. 3.

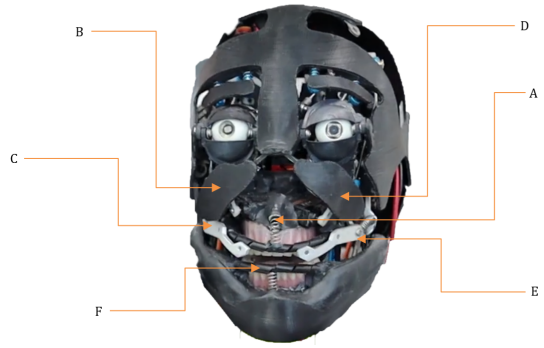

Figure. 3 Left A: Upper lip actuator. B: Right cheek mechanism. C: Right aluminium buccinator actuator. D: Left cheek device. E: Left aluminium buccinator actuator F: Bottom lip mechanism.

The mouth mechanisms are driven by Spektrum DS821 brushless servos with 2.5kg of torque situated in the rear of the robotic head. The servos operate on an Arduino Uno microprocessor with a 12-channel servo shield, which permits an independent power supply at 6v, 3A. A 12V active speaker system is embedded in the base of the neck, and a passive speaker is positioned in the base of the bottom jaw to give a full sound. Aluminium cable wire covered in a silicone wrapping creates manipulatable lips which require little force to actuate using micro servos. The mouth mechanisms are connected to five servos using 100lb tri-bind wire. The gums and teeth are created from acrylic, and the tongue from silicone, Fig. 4 (Left). The tongue is actuated using a length of wire pulled through the centre of the tongue, which raises and lowers the internal mechanism under tension. The robot's skin is attached to the robotic mouth using a series of small, powerful magnets embedded in the silicone skin of the robot, shown Fig 4 (Right).

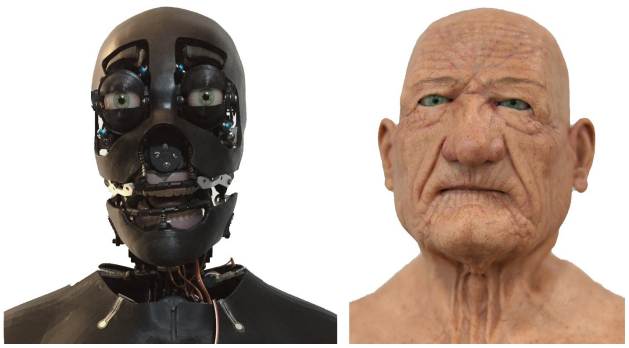

Figure. 4 Complete RHR. Left: Finished exoskeleton. Right: Finished humanoid robot with silicone skin attached.

Medical-grade silicone with a power mesh support embedded into the actuated areas is applied to increase the skin's longevity and strength. This methodology will not restrict the fluidic movement of the skin compared to building up thicker layers of silicone, which may produce unnatural skin and muscular movement. To test the dexterity of the mouth, a series of lip and mouth positions programmed into the Arduino controller provided positional data to ensure the robotic mouth does not operate outside of the parameters of the natural human mouth. This approach is essential for gathering lip and mouth position data before scripting the voice processing application. Failure to comply with this step may result in damage to the silicone skin and internal mechanisms of the robotic mouth by servo motors.

6 Speech Processing to Robotic Mouth Articulation Machine Learning Application

Based on the findings of the literature review, the MPFCC sound analysis method provides effective waveform analysis to extract visemes. Amazon's AI chatbot system 'Lex' implements a DL algorithm which enables a broad scope of intelligent responses. The AI system is configurable with Amazon 'Polly' speech synthesis software for natural speech output. Google's AI 'Dialogflow' provided similar results to Amazon Lex, yet, the embedded speech synthesis software produced a less authentic humanistic voice. The script for the novel voice processing application developed in Arduino (C++) analyses audio data inputs from a PC using a dedicated headphone port. Stripping one end of a stereo headphone cable reveals three wires, red, green and copper. The left and right stereo cables are red and green and attach to an analogue signal port on the Arduino servo shield and the copper wire grounded on the earth pin. To test the strength of the audio signal, open the AnalogReadSerial example and use the serial monitor to view the incoming data field. If the data readings are inconsistent and sporadic or have a low value, i.e. (0-1000), an external amplifier with a noise reduction transistor is required to boost the signal, demonstrated in Figure. 5

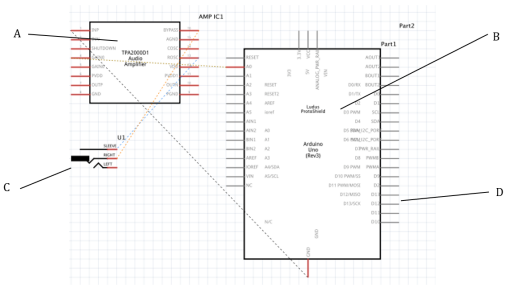

Figure. 5 Arduino Schematic. A: Audio amplifier. B: Arduino Microprocessor with servo shield. C: Audio jack input. D: Digital servo

The optimal signal range for the voice processing application is (0-15000pwm); which is sufficient for acquiring lip position and jaw patterning data. If the data input frequently drops, this may be due to interference from the external amplifier or pc output, applying a 1kw resistor to the ends of the stereo cables reduces noise interference from external electronic devices. The objective of the voice processing system developed in this study is to measure four independent values, firstly, when the audio input reads 0, the mouth and lips close.

This configuration is the default position of the robotic mouth as 0 equates to no sound input from the speech synthesis application. The jaw mechanism operates on the values of (0-10000pwm) this is mapped onto servo motor positions, for example, MouthVal = map (MouthVal, 0, 10000, 0,180). The first data set 0-10000 represents the frequency range of the input audio, and 0-180 is the range of the servomotor. Thus, the higher the amplitude, the wider the mouth opens and the lower the sound input, the less, this is representative of mouth aperture size when talking loud and quiet in humans. Similarly, the frequency and wavelength of the incoming audio transpose into lower jaw pivot speed, this method permits the mouth to perform fast and slow up/down jaw movements in synchronisation with the incoming audio. Secondly, direct speech from Amazon Lex/Polly inputs into the analogue 0 port on the Arduino and transcribed into data over the serial monitor. It is crucial to set a volume level on the system as variation in amplitude effects the accuracy of the system; this approach helps define specific word sounds from the speech synthesis application and creates a stable operating framework for re-application. However, a mid-range audio frequency is optimal as distortion may occur in high levels of audio input which can produce inaccurate lip synchronisation when processed in real-time. The speech synthesis to robotic mouth articulation ML application recognises vowel and consonant patterns in the incoming order using a series of words which inputs into the system as a data stream and measured over the serial monitor.

The objective of this supervised ML method is to find data ranges in the serial monitor that relate to specific vowel and consonant patterns in the incoming audio. The greater the number of words analysed over the serial monitor for their data value, the more accurate the lip synchronisation application will function. This methodology provides a data range for each vowel and constant sound, for example (int L = (3250,3300) ;), meaning all the L sounding words inputted into the voice processing application on average operated between 3250-3300pwm. However, the more words used to train the ML application, the higher the potential for system instability. For example, for int U = (3420, 3530) and int O = (3350,3440), it is crucial to define individual identifiers for specific words, such as 'Hello', which registers between 4536– 4539pwm or create a buffer by reducing or replacing the words used in the system training until the application produces accurate results. The servomotor positions for lip synchronisation operate on 'if' statements, shown in the code sample below.

Setting Mouth Positions for A and I Visemes

int AI =(4570, 4590); // Waveform-range for A and I vowel sounds transcribed into data range int postarg=0; // Target Position for Servo, set at 0 to store value int sensorValue = analogRead(A0); // Audio Input

if (analogRead(LipVal) greaterThan 4560 / lessThan 4600) postarg=AI; // Register incoming data as A and I and send to servos
servo1.write(AI); // Set position for Buccinator actuators
servo2.write(AI+45); // Set position for top lips
servo4.write(AI-23+random 6); // Set low position for Tongue + 6 random degrees for natural movement
servo3.write(AI-13); // Set position for the bottom lip

if (analogRead(LipVal) 4580) postarg=AI; // 4580 is the identifier for the word 'After'as the system struggled correctly to process it over the serial port.

This approach sets the lip positions to match the incoming vowel sounds at the start of each word transcribed from the speech synthesis application. Thus, the greater the sum of words specified and assigned an identifier, the more accurate the ML system operates. However, as microcontrollers have a limited amount of data storage, grouping words that start with the same letters to triangulate vowel and consonant lip positions reduced storage load. Thus, the balance between assigning individual words and grouping words is essential for accurate lip synchronisation using the MFCC method. Thirdly, the tongue positions operate in low, mid and high ranges and not forward and back positions, this function is set using a four-sided phonological vowel chart [31], it is important to set correct tongue positions for O, B and L sounds as the tongue is visible during verbal communication, for instance, if the vowel is (A) the tongue is low. Fourthly, the robotic jaw operates on a similar mapping system, independent of the lip synchronisation but utilising the same analogue 0 input. The jaw functions in two states, 'Moving' when reading incoming data that is greater than 0 and 'Closed' when equal to 0. A speech to mouth articulation delay of 0.1 seconds, smooths mouth speed giving a greater naturalistic and fluidic motion, as annotated in the code below.

Jaws Closed and Open (Talking)

if (analogRead(button)¿0) postarg2==0; //mouth close
if(pos¡postarg2) pos++; // Smoothen Transitions
if(pos¿postarg2) pos–; // Smoothen Transitions
delay (0.1);
int MouthVal = analogRead(A0);
Read Analog Port 0 MouthVal = map (MouthVal, 0, 1023, 0 ,180);
Map Positions postarg2=MouthVal;
servo0.write(postarg2);
Write to Servo

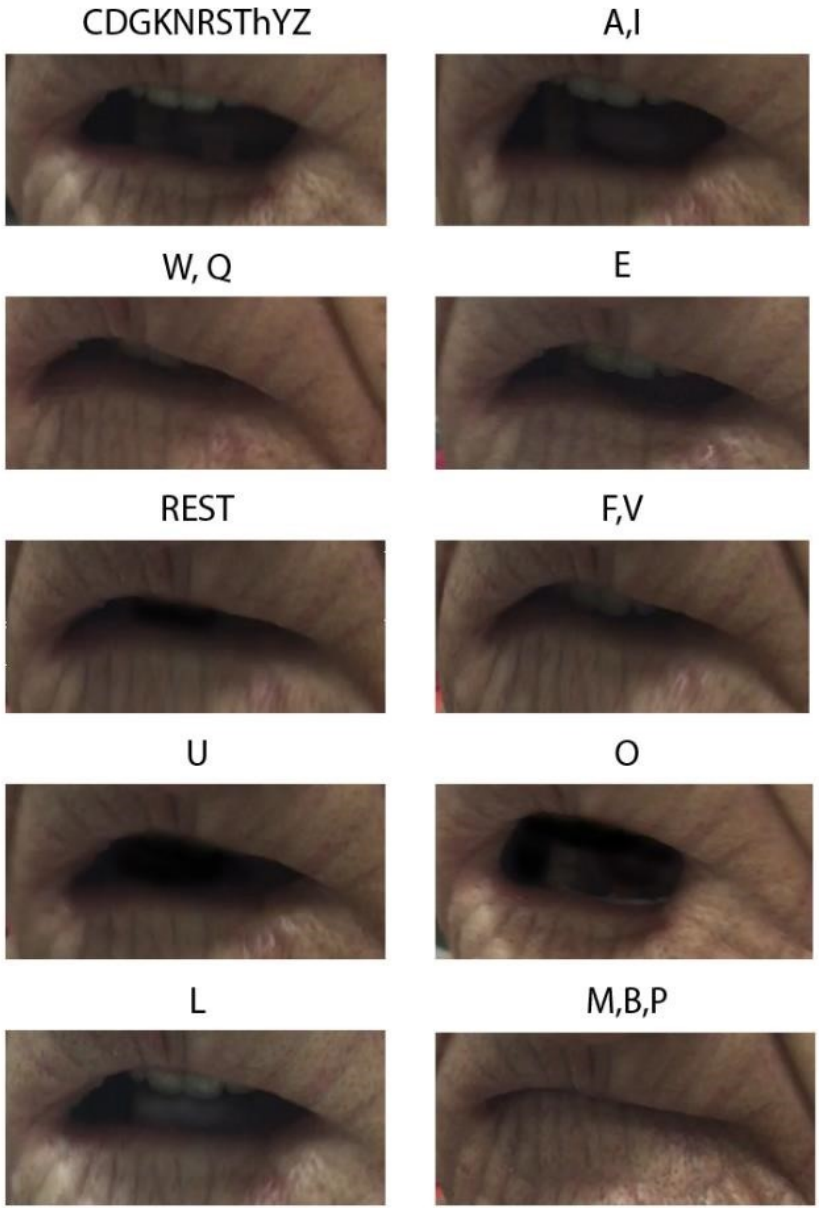

Figure. 6. Robotic Lip and Jaw Shapes to Vowel and Consonant Sounds

Video documentary of the robotic mouth: https://www.youtube.com/watch ?v=iwrRm9Xywas&ab_channel=carlstrathearn

7 Robotic Mouth Evaluation Methodology

The robotic mouth system developed in this study was installed in an RHR named 'Euclid and comparatively analysed against ten of the most advanced contemporary RHRs: Sophia (2015, USA), Fred (2018, UK), Nadine (2008, SG), Junko Chihira (2015, JAP), Geminoid DK (2011, JAP), Bina 48 (2010, USA), Kodomoroid (2014, JAP), AI-DA (2019, UK), Geminoid H1 (2006, JAP) and Alex (2019, RUS) and a human mouth for precision testing and calibration. However, as physical cross-evaluation of the ten robotic systems is impractical due to accessibility issues, this study employs pre-recorded video footage of the robots during verbal communication to evaluate and cross-analyse the RHRs against one another.Therefore, this study employs pre-recorded video footage of the robots during verbal communication to evaluate and cross-analyse the RHRs against one another. Video footage of Euclid taken during a short speech was recorded and added to YouTube and downloaded to reduce potential data-gathering bias by using the same source and evaluation methods as the previous systems. Adobe Premiere Pro video editing software permits the footage to be looped, slowed down, time-matched and spliced.

Assessment of each robot considers the percentage of the correct jaw and lip positions against the natural jaw and lip positions of a human and the time differential between the spoken words and articulation of the robotic mouth. Professional virtual animation lip-synchronisation software 'Adobe Animate' transforms the spoken words into natural vowel and consonant lip positions. The lip synchronisation software consists of a multi-level audio analyser, speech decoder and auto speech to text synchronisation feature for optimum audio/word to lip position splicing and matching. However, the sound quality of each video was highly variable and importing low-quality audio into Adobe Animate produced inaccurate lip-synchronisation due to audio feedback and noise contamination rendered in the original recordings.

Therefore, the words spoken by the robots were transcribed into text and imported into the Amazon Polly speech synthesis application, which provides high-quality MP3 recordings for import into Adobe Animate. Video footage of a human mouth speaking the words of the robots was imported into Adobe Animate and matched/overlaid with the audio and animated lip-synchronisation map to ensure accurate lip positions and measure the time differential between speech and video. All robots evaluated in this study converse in English, apart from the Russian speaking humanoid robot 'Alex'. However, as no footage of Alex speaking English is currently available online, a translator converted Russian into English to calculate the number of syllables, vowels and consonant sounds using standard Russian grammatical principles.

Observational data is collected using an online survey to examine the visual and speech authenticity of the 11 humanoid robots using Likert scales with embedded video samples and deployed to a random sample of 50 anonymous

participants ages 18+.

The following five areas of evaluation formulate the robotic mouth test procedure.

1. Jaw articulation to syllable patterning: The evaluation of the RHRs syllables to jaw motion ratio, measured against the syllables/jaw movement of a human.

2. Speech to mouth articulation differential: The assessment of time difference between the spoken words and mouth movement of the RHRs extracted from video footage.

3. Accurate lip positions during speech: The vowel and consonant lip positions of the RHRs measured against a human.

4. Visual Appearance Authenticity (Aesthetics): The level of authenticity of the RHRs mouths rated on a Likert scale of 1-10. (1: least human, 10: most human)

5. Audible Speech Authenticity: The analysis of the speech synthesis / natural speech transference of the RHRs, measured on a Likert scale of 1-10.

8 Results

The survey for the robotic mouth consisted of 22 quantitative questions set to video and audio based on a recent HRI study [32].The participant sample is 50 random individuals recruited from online social media and forums. The sample consists of 34 (68%) females, and 16/50 (32%) males representing a balanced gender range with an age range between (18-25):11, (26-31):26, (31-40):6 and (40+): 7. The ethnicity's of the sample consisted of 38 (76%) British, 8 (16%) American, 1 (2%) Chinese, 1 (2%) Japanese, 1 (2%) Swedish and 1 (2%) Cypriot, representing a variable range of different ethnic and cultural backgrounds

Statistical Package for Social Sciences (SPSS) analytical software, indicated a moderate-low level of coefficients scoring between 0.6-0.8.The standard deviation ranged between 1.66-2.55 suggesting a high level of dispersion and confidence set at 95% indicates a low margin of error and a low variance ranging between S2:2.7 - S2:6.5, these highly variable results are indicative of the inconsistencies in the reliability of human perception in determining human-likeness, shown in table 1 and 2.

Table.1. Robotic Mouth Questionnaire Statistical Results

Question	**Avg**	**Mode**	**Coeff**	**Std**	**Var**
Q1 Sophia Aesthetics	4.8	5	0.63a	2.27	S2:5.1
Q2 Sophia Speech	5.1	5	0.71a	1.86	S2:3.4
Q3 Fred Aesthetics	6.3	6	0.63a	2.33	S2:5.4
Q4 Fred Speech	6.9	7	0.74a	2.24	S2:5.0
Q5 Nadine Aesthetics	7.1	7	0.73a	2.23	S2:5.0
Q6 Nadine Speech	7.3	7	0.62a	2.05	S2:4.2
Q7 Junko Aesthetics	6.2	6	0.65a	2.05	S2:4.2
Q8 Junko Speech	6.8	7	0.71a	2.17	S2:4.7
Q09 Geminoid DK Aesthetics	7.3	7	0.63a	2.97	S2:4.8
Q10 Geminoid DK	7.4	7	0.68a	2.21	S2:4.3
Q11 Bina 48 Aesthetics	5.8	6	0.8a	1.99	S2:3.6
Q12 Bina 48 Speech	6.2	6	0.68a	1.90	S2:4.9
Q13 Kodomoroid Aesthetics	3.8	4	0.74a	2.06	S2:3.9
Q14 Kodomoroid Speech	5.1	5	0.66a	2.00	S2:3.6
Q15 AI-DA Aesthetics	6.9	7	0.68a	2.09	S2:4.2
Q16 AI-DA Speech	6	6	0.71a	2.12	S2:4.0
Q17 Geminoid H1 Aesthetics	7.2	7	0.69a	2.35	S2:4.5
Q18 Geminoid H1 Speech	6.8	7	0.77a	2.01	S2:5.5
Q19 Alex Aesthetics	6	6	0.68a	2.55	S2:4.0
Q20 Alex Speech	7.3	7	0.74a	2.13	S2:6.5
Q21 Euclid Aesthetics	7.1	7	0.70a	1.82	S2:2.5
Q22 Euclid Speech	7.1	7	0.72a	1.66	S2:2.7

Table.2. Robotic Mouth Video/Speech Test Results

Robot Name	**Class**	**Video Ref**	**Words Eval**	**Jaw Acc (Syllable)**	**Diff (s)**	**Lip Sync Accuracy**	**Visual Auth**	**Speech Auth**
Sophia	WOZ, AI, SS	(A)	19	Human (27) Robot (20) Err (-25.9%) Acc (74.1%)	0.2s	Human (75) Robot (53) Err (-29.3%) Acc (70.7%)	5/10	5/10
Fred	WOZ	(B)	12	Human (18) Robot (12) Err (-33.3%) Acc (66.7%)	0.1s	Human (48) Robot (5) Err (-89.5%) Acc (10.5%)	6/10	7/10
Nadine	AI, SS	(C)	16	Human (24) Robot (20) Err (-25.9%) Acc (74.1%)	0.6s	Human (55) Robot (12) Err (-78.1%) Acc (21.9%)	7/10	7/10
Junko Chihira	AI, SS	(D)	17	Human (27) Robot (22) Err (-18.5%) Acc (81.5%)	0.1s	Human (85) Robot (16) Err (-81.1%) Acc (18.9%)	6/10	7/10
Geminoid DK	WOZ	(E)	14	Human (24) Robot (13) Err (-45.8%) Acc (54.2%)	0.5s	Human (53) Robot (8) Err (-84.9%) Acc (15.1%)	7/10	7/10
Bina 48	AI, SS	(F)	19	Human (23) Robot (15) Err (-34.7%) Acc (65.3%)	0.3s	Human (59) Robot (32) Err (-45.7%) Acc (54.3%)	6/10	6/10
Kodomoroid	SS	(G)	25	Human (32) Robot (12) Err (-62.5%) Acc (37.5%)	0.5s	Human (81) Robot (10) Err (-87.6%) Acc (12.4%)	4/10	5/10
AI-DA	WOZ	(H)	20	Human (29) Robot (26) Err (-10.3%) Acc (89.7%)	0.1s	Human (71) Robot (23) Err (-67.6%) Acc (32.4%)	7/10	6/10
Geminoid H1	WOZ	(I)	18	Human (25) Robot (18) Err (-28%) Acc (72%)	0.2s	Human (62) Robot (11) Err (-82.2%) Acc (17.8%)	7/10	7/10
Alex	WOZ	(J)	10	Human (29) Robot (12) Err (-58.6%) Acc (41.4%)	0.0s	Human (108) Robot (22) Err (-79.6%) Acc (20.4%)	6/10	7/10
Euclid	AI, SS	(K)	17	Human (24) Robot (19) Err (-20.8%) Acc (79.2%)	0.1s	Human (60) Robot (52) Err (-13.3%) Acc (86.7%)	7/10	7/10

9 Analysis of Robotic Mouth Test

Figure. 7 RHR Sophia [33]

1. Sophia ranked as having one of the most authentic robotic mouths in the categories of jaw actuation 74.1% (4th highest) and lip synchronisation 70.7% (2nd highest). Sophia has a mouth differential of 0.2 seconds, suggesting that the translation of speech to mouth articulation is high. However, Sophia combines robotic AI, pre-scripted responses and human control to speech synthesis for decision making; which makes distinguishing which of the three modes were functioning during the video-data impossible to verify. Although this multi-modal approach does not alter the time differential between the speech synthesis application and mouth articulation system, it makes accurately categorising the AI system for comparative analysis problematic. Sophia scored an average 5/10, in the lifelike appearance category and 5/10, for speech authenticity, which was below average for this dataset.

Figure. 8 RHR Fred [34]

2. Fred graded 2.7% above the overall average for precision jaw actuation scoring a 66.7% accuracy rating (6th highest). However, lip synchronisation efficiency was highly insufficient, gauging a 10.5% accuracy rating which is the lowest in the dataset, falling 22.5% below the overall average lip-synchronisation score. Nevertheless, the responsivity of the speech synthesis application to the mouth articulation system was adequate scoring a 0.1s differential rating, indicating that the automated lip-synchronisation system is exceedingly inaccurate but highly responsive to natural speech input. Fred rated 6/10, in appearance, which is average for this subset and 7/10, in speech authenticity, which is the highest frequency in the speech category. These results indicate that Fred has an average or slightly above average rating in all robotic mouth evaluation categories, except lip-synchronisation, which scored significantly lower than the overall average.

Figure. 9 RHR Nadine [35]

3. Nadine scored a 41.7% accuracy rating in the jaw to syllable test falling 19.3% below the overall average, and a 0.6-second mouth to speech differential, ranking lowest in the data set. Similarly, Nadine achieved a 21.9% lip-synchronisation accuracy grading, 11.1% below the overall average score. When comparatively analysed against the three other autonomous robots in the AI and speech synthesis category Nadine ranked (4/4) in the jaw to syllable efficiency test and (3/4) in lip-synchronisation accuracy. However, Nadine is the second oldest robot in the dataset, developed in 2008, therefore, the robot implements outdated systems, mechanical design and electronic actuators. Thus, the results derived from this study are indicative of the limitations of AI and speech processing applications and servomotor accuracy from over a decade ago. Nadine ranked 7/10 in natural mouth appearance and 7/10 in speech authenticity. Therefore, this study concludes that the appearance and speech of Nadine's robotic mouth are substantially more authentic and accurate than the AI and NLP systems, which functioned significantly below the average accuracy rating in the dataset.

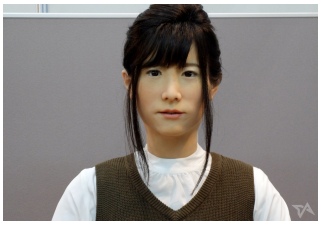

Figure. 10 RHR Junko Chihira [36]

4. Junko Chihira ranked 81.5% (2nd highest in the dataset) fidelity rating in the jaw to syllable synchronisation group. The robot ranked highest for jaw accuracy in the AI and speech synthesis subset, achieving 2.3% greater efficiency than Euclid in second place. Furthermore, the robot displayed a 0.1s-time differential between mouth articulation and speech synthesis output. However, despite the impressive precision of the jaw to syllable patterning system, Junko produced an 18.9% accuracy rating in the lip-synchronisation test, which is the lowest in the AI and speech synthesis subset.

Interestingly, on review of the robot's lip synchronisation data and footage, the actuation of the robotic lips is very subtle, and the emphasis appears not to be on correct lip position but rather a random generation of lip movement during speech. Therefore, it is highly likely that the robot does not implement lip-synchronisation technologies but rather emulates human lip movement using a random sequence generator to control servomotor positions. Junko scored an average 6/10 in authentic mouth appearance and 7/10 in speech authenticity. Junko (AI/speech synthesis) scored highly across all categories expect lip synchronisation, with accuracy readings proximal to Fred's results (human input/natural voice processing). Therefore, the random sequence generator of Junko Chihria's robotic mouth operated with a similar level of accuracy as the automated lip-synchronisation to live speech system of Fred.

Figure. 11 RHR Geminoid DK [37]

5. Geminoid DK rated 52.5% in jaw accuracy during the syllable synchronisation examination, falling short of the average by 11.8% with a 0.5-second voice to actuation processing speed. Similarly, the robot rated low in the lip synchronisation category achieving only a 15.1% precision rating. Geminoid DK ranked (3rd/5) in the human input and human speech processing category showing a high level of inconsistency between the jaw and lip synchronisation performance categories. This result is indicative of the inaccuracies of implementing human input and human voice processing. However, Geminoid DK rated 7/10 in appearance and 7/10 in speech authenticity, which is similar to other systems evaluated in the data set that implement natural speech output methods. Therefore, according to the results of this study, Geminoid DK operates with poor functionality but a high level of aesthetical realism.

Figure. 12 RHR Bina 48 [38]

6. Bina 48 achieved a 65.3% precision score in the jaw actuation to syllable test, operating with an average 0.3-speed differential. The RHR also accomplished a 54.3% lip-synchronisation accuracy rating, well above the 33% average. It is significant to note that Hanson Robotics created both Bina 48 and Sophia, as this is observable in the data gathered during the experiment. Both robots scored highly and within similar ranges in the tests that examined jaw and lip-synchronisation accuracy. The progression of the robotic mouth and speech processing system from Bina 48 created in 2010 to Sophia in 2015 is compelling for reviewing the rate of development for AI and NLP in RHRs. Bina 48 graded 6/10 for aesthetical accuracy and 6/10 in vocal realism. This observation is intriguing as it suggests that although Hanson Robotics has considerably progressed the mouth articulation, AI and NLP speed of their RHRs, there is a noticeable reduction in human-like speech synthesis and humanistic aesthetical quality.

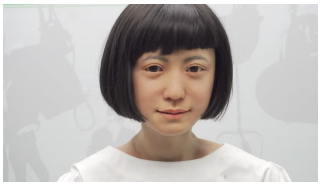

Figure. 13 RHR Kodomoroid [39]

7. Kodomoroid rated poorly across all categories of the mouth and lip synchronisation examination scoring 37.5% in jaw accuracy with a 0.5-second speech to mouth movement differential and 12.4% in lip synchronisation precision. Kodomoroid is a recent RHR developed in 2014. However, Kodmoroid is configured to speak in Japanese, which may account for inaccuracies in speech processing to mouth articulation during English translation. Nevertheless, this does not account for delayed speech to mouth processing speed. Similarly, Kodomoroid ranked 4/10 in aesthetical appearance, which is the lowest score in the dataset and 5/10, just below average in speech authenticity.

Figure. 14 RHR AI-DA [40]

8. AI-DA achieved the highest overall accuracy score of 89.7% in the jaw to syllable actuation test with a 0.1-second speech to mouth articulation processing speed. However, the robot rated significantly lower in the lip-synchronisation category falling short of the 33% average at 32.4%. This result is indicative of the implementation of human voice processing, and human input as the jaw actuation to speech patterning system rated highly, and the lip-synchronisation scored low. As discussed in the literature review, this result is a commonality when deciphering real-time human speech into lip-articulation.

It is vital to state that Fred and AI-DA are creations of the same company 'Engineered Arts' and it is highly likely that they implement similar lip-synchronisation systems, reflected in the results of the examination. AI-DA rated 7/10 in aesthetical appearance and 6/10 in speech authenticity, which is the lowest score in the human input/human voice processing subset for vocal realism. However, these poor outcomes may be a result of the human operator as their vocal performance was highly robotic. Therefore, although AI-DA achieved the highest score in jaw actuation accuracy and above average in appearance, the low ratings in speech authenticity and lip-synchronisation significantly reduce the overall perceptual realism of the RHR.

Figure. 15 RHR Geminoid H1 [41]

9. Geminoid H1 indexed a 72% jaw actuation to syllable accuracy rating (5th highest in the data set). However, the robot achieved a low precision rating of 17.8% in the speech to lip-synchronisation test (4th lowest). Conversely, Geminoid HI scored highly in aesthetical appearance (7/10) and naturalistic speech (7/10). Geminoid H1 ranked on average as the highest scoring robot in the natural human input and human speech output category. This result is particularly intriguing as Geminoid H1 is the oldest robot in the dataset created in 2006 and implements outdated NLP and analogue servos.

This consideration is significant as Kodomoroid, Geminoid H1 and Geminoid DK are creations of the same company 'Hiroshi Ishiguro Laboratory'. However, according to the results of this study, the accuracy and aesthetical authenticity between Geminoid H1: 2006, Geminoid DK: 2011 and Kodomoroid: 2014, has significantly decreased over time.

Figure. 16 RHR Alex [42]

10. Alex graded below average in jaw actuation accuracy achieving 41.1% (22.6% below average) and in lip-synchronisation 20.4% (12.6% below average) which is indicative of the inaccuracies of implementing human control and human speech input (WOZ). However, there was no time differential between the verbal commands and mouth movement of Alex. Thus, it is likely that Alex's speech processing to mouth articulation system operates at a rate proximal to human speech and mouth movement. Alex scored an average 6/10 rating in aesthetical accuracy and 7/10 for speech authenticity.However, as the RHR only speaks in Russian, participants were asked to judge how humanistic the voice sounded rather than understanding the language. This approach produced results within the scope of other findings in the dataset. Therefore, the evaluation of Alex's speech authenticity was not affected by language. Alex 2019 and AI-DA 2019 are the latest RHRs in the dataset, excluding Euclid. However, both robots implement the WOZ approach in place of NLP, machine vision and AI

Figure. 17 RHR Euclid

11. Euclid rated 86.7% (1st highest) in lip synchronisation accuracy and 79.2% (3rd highest) in jaw action to syllable accuracy; this figure is slightly lower than predicted during the initial testing phase before adding lip synchronisation into the programming framework. Therefore, it is likely that the additional system load of the lip-synchronisation system increased processing time or produced acute electronic/audio feedback that affected the accuracy of the jaw component.

However, the additional load did not affect the scripted speech to mouth differential of 0.1s. A potential solution to overcoming this issue in the future is to run the jaw and lip-synchronisation features on separate microcontrollers using the same audio input from the speech synthesis application; this approach may increase stability by sharing the system load between two controllers. The robotic tongue element of Euclid did not appear to influence the perceptual authenticity of the robot during operation. However, as the robot was restricted to evaluation by video footage, this feature was not apparent to the viewer during recording. Euclid achieved 7/10 for aesthetical authenticity, which is the highest frequency in the subset and 7/10 for realistic speech. These results are particularly encouraging as Euclid achieved, on average, high ratings across all five examination categories of the dataset and higher in functional accuracy than [11], recent robotic mouth prototype for RHRs.

10 Conclusion

This paper highlights a significant bottleneck in RHR engineering and produced a robotic mouth and speech processing application which operated with a high levels of functional and aesthetical realism/accuracy. The novel robotic mouth system permits the widening/contracting and raising/lowering of the corners of the RHR's silicone lips to emulate human lip patterns and shapes (visemes) using a custom set of buccinator actuators. A custom robotic lip synchronisation application controls this component, using an ML AI approach which clusters the PWM frequencies of similar-sounding words from a set of pre-defined rules and example data.

The robotic tongue emulates the up and down positions of the human tongue during the pronunciation of vowel and consonant sounds, based on the four-sided phonological vowel chart [31]. However, as the internal components of the robotic mouth were problematic to capture using the video-based evaluation methodology for the robotic mouth calibration and comparative examination, the influence of accurately emulating human tongue movements during verbal communication for increasing human-likeness and enhancing language understanding (mouth reading) in HRI is difficult to verify. The robotic mouth system was tested against 10 of the most technologically advanced and realistic RHRs using a pre-recorded video analysis method and sophisticated animation software for generating visemes for lip synchronisation. The robotic mouth achieved the highest rating for lip-synchronisation to speech accuracy in the dataset. Robotic jaw actuation to syllable accuracy was slightly reduced when running the system in conjunction with the lip and tongue subroutines of the robotic mouth application. This outcome was not expected as during the initial testing of the robotic mouth prototype before the addition of the lip/tongue synchronisation element, the system achieved a significantly higher accuracy rating.

Despite the limitations of the Arduino microprocessor in handling large sums of data input/output, the robotic jaw achieved the 3rd highest rating in the jaw to syllable accuracy examination and 2nd in the speech to mouth actuation differential test. On average, the aesthetical quality and speech authenticity of Euclid achieved the highest ratings for human-likeness in the dataset, ranking joint first. It is essential to account that the developmental process using CAD, CGI animation rigs and 3D printing methods proved essential in the creation of the robotic mouth system. Finally, the robotic mouth developed in this study operates with a higher degree of precision and human-likeness using the novel buccinator actuator system, 3D printed robotic mouth design and speech synthesis to robotic mouth actuation application, ranking top for lip synchronisation accuracy in the dataset. As per the outcomes of this robotic mouth calibration experiment, the implementation of the buccinator actuator system, robotic mouth model and adjoining software in future RHR design will increase the accuracy and authenticity of RHR mouth design.

11 Declarations

Funding (This project was funded by Staffordshire University as a PhD Scholarship)

Conflicts of interest/Competing interests (The authors declare no conflict or competing interests in the publication of this research)

Ethics approval (Ethical approval for this project was obtained via the research ethics committee of Staffordshire university on 15.05.19)

Consent to participate (All participants in this study consented to participation in accordance with GDPR rules)

Consent for publication (The authors consent to this research being published)

Availability of data and material (All data is included in the text)

Code availability (The code for this project is presented in the text)

Authors' contributions (Dr Carl Strathean is the primary investigator and Prof. Eunice Minhua Ma was his supervisor)

12 References

[1] Dang. S, Artificial Intelligence in Humanoid Robots, Retrieved www.forbes.com/sites/cognitiveworld/2019/02/25/artificial-intelligence-inhumanoid-robots/87a95fb24c72, (2019). Accessed 17.11.19

[2] Urbi. J, The Complicated Truth About Sophia the Robot, An Almost Human Robot or a PR stunt. Retrieved: www.cnbc.com/2018/06/05/hanson-roboticssophia-the-robot-pr-stunt-artificial-intelligence.html. (2018) Accessed: 17.11.19

[3] Rosenthal-von. M, Kramer. N, Maderwald. S, Bran .M, Grabenhorst. F, NeuralMechanisms for Accepting and Rejecting Artificial Social Patterners in the Uncanny Valley, Journal of Neuroscience, Vol.39, Issue 33, pp.6555-6570. DOI:10.1523/JNEUROSCI.2956-18.2019, (2019)

[4] Strathearn. C, Ma. M. Modelling User Preference for Embodied Artificial Intelligence and Appearance in Realistic Humanoid Robots, Informatics:, 7, 28.DOI:doi.org/10.3390/informatics7030028 (2019)

[5] Avril, T The art of speech-reading: One angry eagle's fan helps us learn the intricacies behind reading lips. Lip-reading. pp.47-52. (2019)

[6] Klein. J When Mismatched Voices and Lips Make Your Brain Play Tricks. Retrieved:www.nytimes.com/2017/02/21/science/lip-reading-mcgurk-effect.html. (2017) Accessed: 17.12.20

[7] Llorach .G, Evans .A, Blat .J, Grimm. G, Homann .V, Web-Based Live SpeechDrive Lip-Sync, 8th International Conference on Games and Virtual Worlds for Serious Applications. DOI: 10.1109/VS-GAMES.2016.759038 (2016)

[8] Cintas, R, Cid. R, Manso, L.J., Calderita, L., Sanchez, A., & Núñez, P.A. A real-time synchronisation algorithm between the Text-To-Speech (TTS) system and Robot Mouth for Social Robotic Applications. 19th Symposium in Robot-Human Communication Interaction. DOI:10 .1109/ROMAN.2010.5598656 (2011)

[9] Engel. O, Making the Machine Believable: Wizard of Oz-ing AI Applications, Retrieved from: https://uxdesign.cc/making-the-machine-believable-wizard-of-oz-ing-aiapplications-293cfbb0f244. (2009) Accessed 17.11.19

[10] Jain. P. Rathee. M, Anatomy, Head and Neck, Stylopharyngeus Muscles. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK547719/.(2019) Accessed: 17.11.19

[11] Arias. E, Encalada, P, Tigre. F, Granizo. C, Gordon. C, Garcia. M. A ConvNetBased Approach Applied to the Gesticulation Control of a Social Robot. The Interactional Conference on Advances in Emerging Trends and Technologies. pp.186-195. (2019)

[12] Mori, M. The uncanny valley. Energy, Vol. 7, pp.33-35. (1970)

[13] Foley. C. Have we crossed the Uncanny Valley? Retrieved from: https:/ /medium.com/vuto ken/have-we-crossed-the-uncanny-valley-b4512bc6 4b9c.(2018) Accessed: 27.03.19

[14] Burleigh, T. Schoenherr, J. Lacroix, G. Does the uncanny valley exist? An empirical test of the relationship between eeriness and the human likeness of digitally created faces, Computers in Human Behaviour, 29 (3), pp. 759-771. https://doi.org/10.1 016/j.chb.2012.11.021. (2013)

[15] Wang, S, Lilienfeld, S & Rochat, P. The Uncanny Valley: Existence and Explanations, Review of General Psychology, APA Press, 12(4), pp.393-407. https://doi.org/10.1037/gpr0000056 (2015)

[16] Guizzo, E. Who is Afraid of the Uncanny Valley? Retrieved from: https://spe ctrum.ieee.org/automaton/robotics/humanoids/040210-who-is-afr aid-of-the-uncannyvalley. (2010) Accessed: 12.02.19

[17] Schwind, V. & Jäger, S. The Uncanny Valley and the Importance of Eye Contact. In: Ziegler, J. (Hrsg.), i-com: Vol. 15, No. 1. Berlin: De Gruyter. (S. 93–104). DOI: 10.1515/icom-2016-0001. (2016)

[18] Grimshaw.M, Tinwell .A, Nabi .A.D, Williams. A, Facial expression of emotion and perception of the Uncanny Valley in virtual characters. Journal of Computers in Human Behaviour, Volume 27, Issue 2, March 2011, pp 741-749 (2011)

[19] Tromp. J. Bullock. A Steed. A Sadagic. A Slater. M, Frécon. E. Small-group behaviour experiments in the COVEN project. IEEE Computer Graphics and Applications, 18 (6) pp. 53-63. DOI: 10.1109/38.734980. (1998)

[20] Nass. C, Isbister. K and Lee. E. J. Truth is a beauty: Researching conversational agents. Embodied conversational agents, MIT Press, pp. 374-402. (2000)

[21] Garau, M, Slater. M, Vinayagamoorhty. V, Brogni. A Steed. A. and Sasse, M. A. The Impact of Avatar Realism and Eye Gaze Control on Perceived Quality of Communication in a Shared Immersive Virtual Environment. CHI 2003: Proceedings of the SIGCHI conference on Human factors in computing

systems, 05(1) pp. 529-536.(2003)

[22] McGurk H, MacDonald J. Hearing lips and seeing voices. Nature 264, 746– 748 DOI:10.1038/264746a0, (1976)

[23] Ciechanowski. S, Przegalinska. L, Aleksandra, Magnuski, Mikołaj & Gloor, P. In the Shades of the Uncanny Valley: An Experimental Study of Human-Chatbot Interaction. Future Generation Computer Systems. 10.1016/j.future.2018.01.055. (2018)

[24] Elden. S.L. Introduction to Natural Language Processing (NLP). Retrieved from: https://towardsdatascience.com/introduction-to-natural-lang uage-processing-nlp323cc007d f3d. 11.11.19. (2019)

[25] Cid Burgos, F, Manso Fernández-Arguelles, L, Calderita, L, Sánchez, A, & Núñez Trujillo, P. Engaging human-to-robot attention using conversational gestures and lip-synchronization. Journal of Physical Agents, 6(1), 3-10. DOI: https://doi.org/10.14198/JoPha.2012.6.1.02. (2012)

[26] Oh, K., Jung, C., Lee, Y., & Kim, S. Real-time lip synchronisation between the text-to-speech (TTS) system and robot mouth. 19th International Symposium in Robot and Human Interactive Communication, pp.620-625. (2010)

[27] Hara. F & Endo. K. Dynamic control of lip-configuration of a mouth robot for Japanese vowels. Robotics and Autonomous Systems. 31. pp.161-169. DOI:10.1016/S0921-8890(99)00105-0.(2000)

[28] Hyung. H, B. Ahn, D. Choi, D. Lee and D. Lee. "Evaluation of a Korean Lip-sync system for an android robot," URAI, Xi'an, pp. 78-82. doi: 10.1109/URAI.2016.7734025 (2016)

[29] Ailm. A & Rashind. S. Some Commonly Used Speech Feature Extraction Algorithms, Intech Open Journal, Submitted: October 4th 2017. DOI: 10.5772/intechopen.80419 (2018)

[30] Jamaludin, A., Chung, J.S. & Zisserman. A. You Said That?: Synthesising Talking Faces from Audio Int J Computer Vision, 127: 1767. https://doi .org/10.1007/s11263-019-01150-y, (2019)

[31] Tun. K. English Phonetics and Phonology for Burmese-Myanmar Speakers. 2nd ed., 4th printing 1993, Cambridge University Press. ISBN 0-521-40718-4. pp 262 (2009)

[32] Bartneck. C, Croft. E & Kulic. D. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived

safety of robots. Int Journal of Social Robotics, 1(1), 71-81. DOI:10.1007/s12369-008-0001-3 (2009).

[33] Hanson. D. SOPHIA THE ROBOT. Retrieved: https://londonspeakerbureau.com/speaker-profile/sophia-the-robot/. (2017), Accessed 18.12.20

[34] Engineered Arts, Fred The Humanoid Robot. Retrieved: https://www.engineeredarts.co.uk/fred-mesmer-promo/. (2019) Accessed 19.12.20

[35] Silverberg. D. Nadine the Robot Is Your New Social Companion. Retrieved: https://www.vice.com/en/article/pgkgzv/nadine-the-robot-is-your-new-social-companion-Nadia-Thallman.(2016) Accessed: 19.12.20

[36] Quigley. J.T. The future or just freaky? Face to face with 3 of Japan's humanoid robots. Retrieved: https://www.techinasia.com/irex-2015-japan-humanoid-robots.(2015) Accessed: 19.12.20

[37] Geminoid DK. Robotics. Retrieved: https://robots.ieee.org/robots/geminoiddk/.(2015) Accessed: 19.12.20

[38] James. M. Seven Days Chats Up Vermont's Most Interesting "Talking Head," Bina48. Retrieved: https://www.sevendaysvt.com/vermont/seven-days-chats-up-vermonts-most-interesting-talking-head-bina48/Content?oid=2266351. (2013) Accessed: 19.12.20

[39] Science and Technology. Robots, 2017 at the Science Museum in London, United Kingdom Retrieved: https://wsimag.com/scienceandtechnology/26336-robots. (2017), Accessed: 19.12.20

[40] Block. I, AI robot Ai-Da presents her original artworks in University of Oxford exhibition. Retrieved: https://www.dezeen.com/2019/06/14/ai-robot-ai-da-artificial-intelligence-art-exhibition/. (2019) Accessed: 19.12.20

[41] Ishiguru. H. Geminoid H1. Retrieved: http://www.geminoid.jp/en/robots.html. (2012), Accessed: 19.12.20

[42] Malewar. A. A humanoid robot Alex maybe the futuristic news anchor. https://www.techexplorist.com/humanoid-robot-alex-futuristic-news-anchor/22439/. (2019) Accessed: 19.02.20

13 Video References (Table. 2)

A. www.youtube.com/watch?v=78-1MlkxyqI. Dur (2.04-2.10) Acc (12.10.19)

B. www.youtube.com/watch?v=0UufMROVIaU. Dur (0.11-0.16) Acc (21.10.19)

C. www.youtube.com/watch?v=cvbJGZf-raY Dur (0.07-0.13) Acc (12.10.19)

D. www.youtube.com/watch?v=65W2Vn6payw Dur (0.00-0.17) Acc (21.10.19)

E. www.youtube.com/watch?v=NSLe7xrP4jQ Dur (0.28-0.32) Acc (03.11.19)

F. www.youtube.com/watch?v=mfcyq7ugbzg. Durr (0.00-0.16) Acc (03.11.19)

G. www.youtube.com/watch?v=BP-jMfiH-PY Dur (0.00-0.16) Acc (03.11.19)

H. www.youtube.com/watch?v=2HiAjQmpr1w. Dur (0.39-1.08) Acc (21.10.19)

I. www.youtube.com/watch?v=2HiAjQmpr1w Dur (0.17-0.25) Acc (21.10.19)

J. www.youtube.com/watch?v=UZ4twA30Wu4 Dur (0.00-0.07) Acc (21.10.19)

K. www.youtube.com/watch?v=DA9i0z-1sR4 Dur (0.10-0.22) Acc (11.11.19)