Facial Expression Control and Generation of Sophia Humanoid Robot

A project of the 2024 Robotics Course of the School of Information Science and Technology (SIST) of ShanghaiTech

https://robotics.shanghaitech.edu.cn/teaching/robotics2024

许 腾, Teng Xu, 2021233339, xuteng@shanghaitech.edu.cn
周涛涛, Taotao Zhou, 2023233227, zhoutt2023@shanghaitech.edu.cn

June 26, 2024

1 Abstract

Our project focuses on improving interaction capabilities in humanoid robots, specifically targeting the facial expression system of Sophia, a sophisticated humanoid developed by Hanson Robotics. Although Sophia is capable of mimicking a broad range of human facial expressions, her control systems lag in real-time, dynamic expression replication, relying heavily on artist-created, predefined expressions. We address these limitations by developing a transformer-based architecture to map ARKit parameters to motor parameters for accurate reproduction of facial expressions from videos or live captures. Additionally, we are establishing a real-time translation pipeline from text to facial expressions, enabling Sophia to converse with dynamically generated expressions. These advancements enhance Sophia's responsiveness in interactions, contributing to the broader field of robotics by improving humanoid robots' interactive capabilities and facilitating more nuanced human-robot communication.

2 Introduction

The intersection of robotics, artificial intelligence, and human-computer interaction has yielded extraordinary developments in recent years, among which humanoid robots stand as a milestone of technological advancement. Sophia, a humanoid robot developed by Hanson Robotics, encapsulates these advancements. Since her debut in 2016, Sophia has been considered as one of the most human-like robots by the public and researchers, symbolizing the potential of the humanoid robots' capabilities to interact with human beings. Her ability to mimic human facial expressions through a sophisticated array of motors embedded beneath a skin-like material called "Frubber" places her at the forefront of interactive robotics. This innovative design allows Sophia to exhibit a range of emotions and reactions, making her one of the most lifelike robots in existence.

Sophia's expressiveness is facilitated by an advanced facial motion system with 33 Degrees of Freedom (DoF). This system enables her to display a spectrum of emotions with remarkable nuance, from joy and sorrow to curiosity and contemplation. Each movement is powered by precisely coordinated actuators that replicate the actions of human facial muscles, such as the zygomaticus major for smiling or the orbicularis oculi for blinking. These actuators presents the robot's mechanical sophistication and represent a significant step towards creating robots capable of genuine human-like interactions.

Despite these advancements, a critical challenge persists: the gap between Sophia's mechanical capabilities and the sophistication of control systems needed for dynamic, real-time expression replication. Currently, Sophia's expressions are primarily pre-defined and created by artists, with her control system limited to executing these preset expressions. This approach, while effective for demonstrating a range of facial movements, falls short of achieving the spontaneous and responsive interaction that characterizes genuine human expression.

Recognizing this gap, our project aims to pioneer a new frontier in humanoid robotics by developing an expression control system for Sophia that leverages cutting-edge AI techniques for real-time, dynamic expression generation and replication. Specifically, we propose to:

- Develop a facial expression generation scheme using Deep Learning and other self-supervised learning methods, enabling Sophia to autonomously generate expression driven by humans in real-time.
- Create a control method capable of reproducing facial expressions from videos or live captures with high fidelity, thus enhancing Sophia's interactive capabilities.
- Establishing a real-time text-expression translation pipeline. Integrating ChatGPT' s AI response text and neural models to translate audio into expressions, this pipeline will allow for real-time conversational capabilities, enabling Sophia to respond with dynamically generated expressions during interactions.

This dual approach seeks not only to augment Sophia's expressive abilities but also to contribute to the broader field of robotics by exploring novel methods for improving the interactive capabilities of robots. By advancing the underlying technology for expression control in humanoid robots, our project endeavors to enhance human-robot interaction, paving the way for robots to engage more naturally and effectively with humans. This initiative holds the promise of transformative impacts across diverse domains, including education, healthcare, and entertainment, where enhanced human-robot interaction can deliver significant benefits.

3 State of the Art

The foundational work by Chuang et al. [CB02] introduces a method for generating realistic facial animations using blendshape interpolation driven by performance data. This approach leverages motion capture or keyframe animation to enhance the naturalness of facial expressions by interpolating between different blendshapes. The methodology involves aligning performance data with a character's facial rig, mapping this data to blendshape weights, and generating smooth transitions between expressions, forming the basis for realistic and expressive facial animations.

To be specific, the paper provides an overview of blendshape interpolation techniques, which involve creating a set of "blendshapes" or key shapes representing different facial expressions (e.g., smile, frown, raised eyebrows). These blendshapes are then linearly interpolated to generate intermediate facial expressions. The paper introduces the concept of performance data, which includes motion capture data captured from human actors or keyframe animation created by animators. Performance data provides realistic motion and expressions that can be used to drive the blendshape interpolation process. The paper describes the proposed methodology for performance-driven facial animation using blendshape interpolation. This involves aligning the performance data with the target character's facial rig, mapping the performance data to blendshape weights, and interpolating between blendshapes to generate smooth and natural facial animations. In our Sophia system, one of our assumption is that we use blendshapes as a proxy to the movements of the actuators in the Sophia's face. We collect our data in the format of blendshapes of Apple's ARKit, and then map them into Sophia's motor movements. This paper gives the fundamental theory of how we might manipulate the blendshapes to create different facial animations initially.

Building on this foundation, Thambiraja et al. [THA+23] delve into personalized speech-driven 3D facial animation with their system, Imitator. This research focuses on the dynamic generation of facial expressions synchronized with speech inputs. By analyzing speech signals and mapping them to facial expression parameters, the system enables real-time animation, enhancing the interactivity and expressiveness of digital avatars.

Shi et al. [SWJD23] further explore motion generation through deep reinforcement learning (DRL), presenting a sophisticated method for producing diverse and high-fidelity human motion sequences. This paper highlights the potential of DRL in controlling and generating complex motion patterns, providing a robust framework for various applications in animation and robotics.

In the realm of mobile manipulation, Xia et al. [XLMM+21] propose ReLMoGen, an innovative approach integrating motion generation with reinforcement learning for mobile manipulation tasks. By defining tasks as time-discrete Partially Observable Markov Decision Processes (POMDPs) and using subgoals for motion generation, this method improves the efficiency and effectiveness of visuomotor control in complex environments.

Specifically, the paper presents a new trajectory learning scheme for a limb exoskeleton robot designed to assist patients with lower limb disabilities. This approach combines Dynamic Movement Primitives (DMP) with Reinforcement Learning (RL) to generate walking motions. The exoskeleton has six degrees of freedom, focusing on the hip and knee and the ankle. A five-point segmented gait planning strategy is employed for trajectory generation, ensuring stability via a Zero Moment

Point margin. The method effectively addresses joint uncertainties and interferences, enhancing motion generation for exoskeleton-assisted walking.

The Deep Reinforcement Learning (DRL) in the study is used to refine the walking trajectories generated by the Dynamic Movement Primitives (DMP) for an exoskeleton robot. RL helps in adapting to uncertainties in joint movements by learning from interactions with the environment to achieve optimal motion. This process involves iteratively improving the policy that dictates the robot's actions based on feedback from the environment, aiming to minimize errors and enhance the stability and efficiency of the exoskeleton's walking patterns.

Shifting the focus to robotics and exoskeletons, Zhang et al. [ZZ22] address the development of walking exoskeleton robots by combining Dynamic Movement Primitives (DMPs) with reinforcement learning to enhance motion generation capabilities. This study aims to optimize the walking patterns of exoskeletons, significantly contributing to the field of rehabilitation robotics by improving the safety and effectiveness of these assistive devices.

Zhang et al. [ZXCC23] investigate multi-objective optimal trajectory planning for robotic arms using deep reinforcement learning. This research addresses the challenges of balancing multiple objectives, such as accuracy, energy consumption, and smoothness of motion. By integrating these factors into the reinforcement learning environment, the study showcases enhanced precision and efficiency in robotic arm movements, pushing the boundaries of what can be achieved in robotic trajectory planning.

Furthermore, Chen et al. [CHL+21] presents a significant advancement in animatronic robotics and facial expression mimicry by developing Eva 2.0, an animatronic robotic face with soft skin and flexible control mechanisms, capable of learning and imitating diverse human facial expressions through a vision-based self-supervised learning framework. Previous works in animatronic robotics have primarily focused on hardware design and pre-programmed facial expressions, as seen in systems like Kismet [BBM+98] and Albert HUBO [OHK+06], which are limited by their reliance on fixed sets of expressions and extensive human effort. Recent studies have introduced more general motion control mechanisms, such as Affetto's motor displacement modeling and XIN-REN's [RH16] facial feature tracking, but these approaches also face limitations in generalization and online inference.

Significant progress in synthetic video generation and character face animation has focused on motion re-targeting, with methods like Face2Face[TZS+16] and X2Face[WSKZ18] transferring motions across different subjects or domains, though these are restricted to digital avatars and require detailed 3D knowledge. Imitation learning research typically centers on manipulation, locomotion, and navigation rather than facial mimicry, posing unique challenges for achieving precise and varied expressions. The authors of this paper propose a novel two-stage learning framework comprising a generative model that synthesizes a robot self-image from normalized human facial landmarks and an inverse model that outputs motor commands from the synthesized image. This framework enables the robot to learn from a single motor babbling dataset without requiring human labels or predefined expression sets, demonstrating accurate and diverse facial mimicry with real-time response capabilities. Comprehensive evaluations show that this method outperforms nearest-neighbor search-based algorithms and direct mapping methods, effectively generalizing to diverse human subjects and expressions. The contributions of this work include an advanced animatronic robotic face design, a vision-based self-supervised learning framework, and a comprehensive human facial expression dataset, paving the way for more natural and engaging human-robot interactions. Future research directions include incorporating additional modalities such as speech signals to enhance interactive social behaviors in robots.

Specifically, it addresses similar challenges as our project in facial expression control and dynamic, real-time expression replication. Our project aims to enhance the interaction capabilities of the Sophia Robot by bridging the gap between its advanced mechanical design and the sophistication of its control systems. The approach taken in the paper, which utilizes a vision-based self-supervised learning framework to enable real-time, adaptive facial mimicry, aligns closely with our goal of developing a more responsive and nuanced control system for Sophia. By leveraging techniques such as Deep Reinforcement Learning and self-supervised learning, we can draw on the methodologies and findings presented in the paper to inform our development of a facial expression generation scheme. The dual-stage learning framework proposed in the paper—comprising a generative model and an inverse model—offers a viable pathway for creating a control method capable of high-fidelity expression replication from videos or live captures. Thus, the insights and results from this paper provide a foundational basis for addressing the critical issues identified in our project, paving the way for more advanced and effective facial expression control systems in humanoid robots like Sophia.

Each of these studies contributes to the broader field of robotics and animation by advancing our

understanding and capabilities in motion generation, facial animation, and reinforcement learning. The insights gained from these works provide a solid foundation for future research and development in creating more realistic, efficient, and interactive robotic systems.

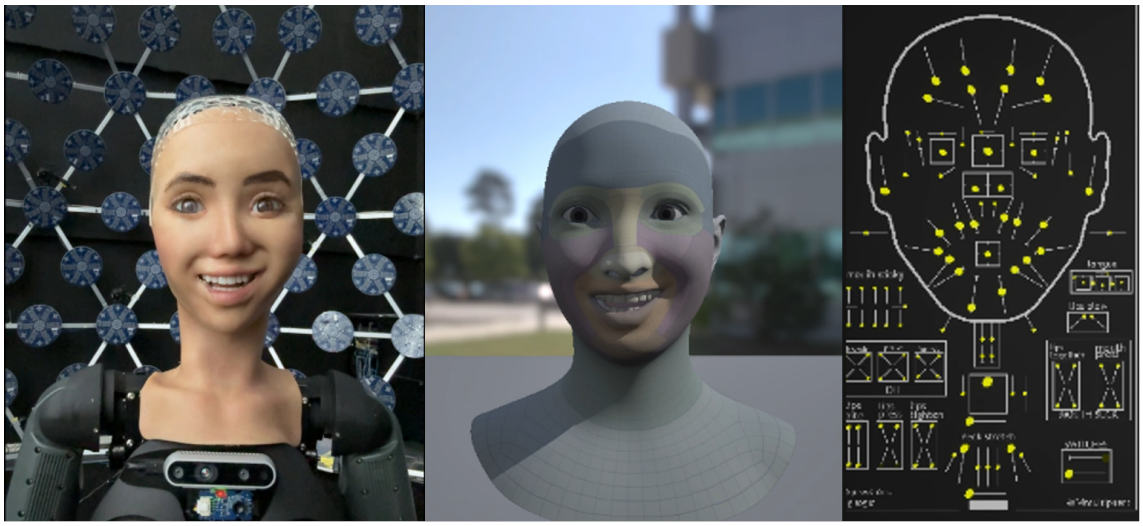
Also, there are several relevant ros packages to our project controling Sophia's facial movement. One way we would like to explore the Sophia's facial expression control is to use the **openai_ros** package to do deep reinforcement learning. The **openai_ros** package serves as a crucial bridge between ROS and OpenAI's Gym interface, facilitating the development and training of AI models within the robotics context. This integration package allows us to apply state-of-the-art reinforcement learning algorithms and techniques, provided by OpenAI Gym, directly to physical robots or simulations in a ROS environment. The **openai_ros** package abstracts the complexities involved in connecting ROS with the Gym interface, thereby enabling developers to focus on designing, training, and evaluating their models without worrying about the underlying communication layer. By leveraging this package, developers can simulate real-world scenarios within a controlled environment, test their algorithms with various robotic sensors and actuators, and iterate rapidly through the development cycle. This accelerates the process of creating intelligent behaviors for robots, making it an invaluable tool for robotics research and development.

For the project aimed at enhancing the expression control of Sophia Robot, the **openai_ros** package presents a powerful platform for training and evaluating deep reinforcement learning or self-supervised learning models. Utilizing this package, we can simulate Sophia's facial expression mechanisms and the associated control system within a ROS environment, allowing for extensive experimentation and optimization. Furthermore, the ability of **openai_ros** to integrate with real-world data and feedback loops enables the trained models to be finely tuned for accuracy and responsiveness.

Besides, we can also leverage ROS RL packages to enhance the control of Sophia's facial expressions. By setting up a ROS environment, we'll simulate interactions with Sophia's control system and create a structured learning environment. Here, actions (altering facial expressions), observations (feedback from expression motors or sensors), and rewards (expression accuracy) will guide the development of a reinforcement learning model. This model will iteratively learn to optimize Sophia's expressions for increased accuracy and responsiveness, benefitting from the rich resources and examples provided in ROS RL documentation.

4 System Description

4.1 Problem Statement



(1) Captured Sophia Image (2) MetaHuman Animation (3) Parameter Response

Figure 1: (1) shows the captured Sophia image $\mathbf{x}$, which is driven by motor parameters $\mathbf{m}$ through the forward kinematics process $\mathcal{F}$. (2) shows the animated results of facial expression detector MetaHuman $\mathcal{N}$. (3) is the visualization of the distribution and magnitude of blendshape parameters $\mathbf{b}$.

Given an input image of a real human, we want to achieve a robust expression transfer from

the real human to our humanoid robot Sophia.

To digitalize expression and the process of transfer, we use a reliable commercial facial detector $\mathcal{N}$ (in our case, it is MetaHuman) to obtain the blendshape parameters $\mathbf{b}$ from captured image $\mathbf{x}$. Since Sophia is fully controlled by motors. Her expression is described by motor parameters $\mathbf{m}$. Taken motor parameters as input, we can drive Sophia by a forward kinematics process $\mathcal{F}$, then use a camera to capture the expression on Sophia. To simplify, we denote the output of $\mathcal{F}$ is the captured expression image $\mathbf{x}$ driven by the motor parameters $\mathbf{m}$. The core problem lies in solving the mapping from the blendshape params to the motor parameters. We denote the mapping to be learned as Φ.

We can use the following equations to formulate our problem:

$$\mathcal{N}(\mathbf{x}) = \mathbf{b} \tag{1}$$

$$\Phi(\mathbf{b}) = \mathbf{m} \tag{2}$$

$$\mathcal{F}(\mathbf{m}) = \mathbf{x} \tag{3}$$

Figure 2: In this figure, we show our captured expression dataset. It contains the captured RGB frames and the corresponding blendshape parameters detected by MetaHuman. We visualize the blendshape parameters in the Unreal Engine, as shown in the second row.

4.2 Algorithm

We developed an facial expression generation scheme for the humanoid robot Sophia based on a deep learning method. Furthermore, we create a pipeline for automatic expression generation from text, while Sophia is chatting with humans, she can talk with various generated facial expressions.

4.2.1 Offline expression transfer using a recorded dataset

In this stage, we adopt a Transformer network as the mapping Φ. It takes in the blendshape parameters $\mathbf{m}$ and predicts the motor parameters $\mathbf{m}$.

To perform the training process, we first captured an expression dataset, as shown in 2. The dataset has ground truth motor parameters and pair-wise blendshape parameters detected by MetaHuman. We use about 50 preset animation clips of Sophia. The total lengths of the clips amount to around 1 hour. Plus that, we randomly sampled Sophia's motor space to generate more expressions. The distribution of the sampled motor parameters is shown in 3. Then we put an iPad in front of the humanoid robot Sophia to capture the animation videos in 30 fps.

The first problem we have overcome is the synchronization of blendshape parameters and motor parameters. Since there are accumulated errors in the timestamp and the motor parameters. Every 10 seconds, we perform a synchronization using the "Eye Blink" expression as an obvious signal.

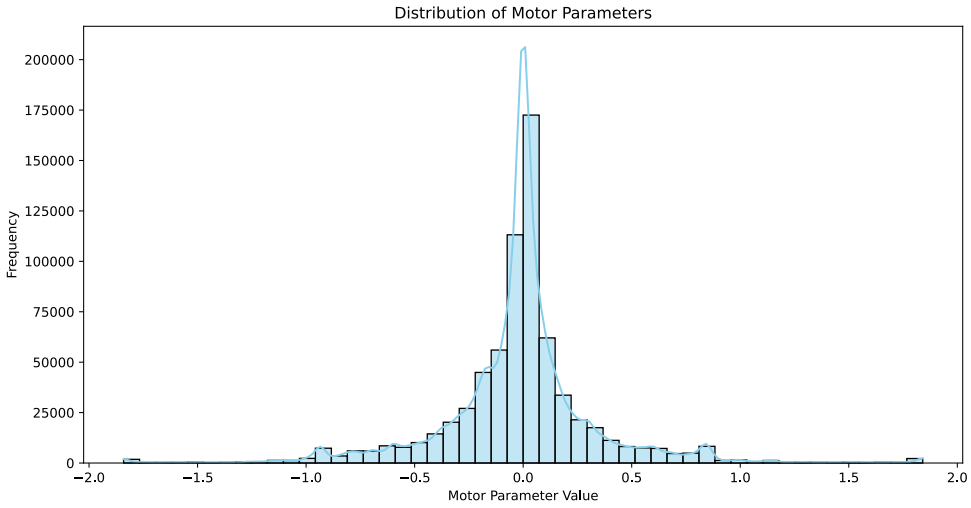


Figure 3: The distribution of training motor parameters.

Once in the parameter graph, we detect the "Eye Blink" signal, we can automatically align the motor parameters and blendshape parameters. The error of our synchronization is around 1-2 frames. Once we have successfully synchronized the data, we can capture the ground-truth dataset of Sophia's expression in a long sequence without any stops.

Next, given the recorded dataset, we trained an Transformer network, with embedding size 64, to learn the mapping between blendshape parameters to the motor parameters. We split the dataset into a training set and a testing set by a ratio 7:3. The network's forward process can be formulated as:

$$\mathbf{m}_{pred} = \Phi(\mathbf{b})$$

The loss function we use is:

$$L_1 = MSE(\Phi(\mathbf{b}), \mathbf{m}_{gt})$$

The loss function L_1 directly supervised the predicted motor parameters using the captured ground truth dataset.

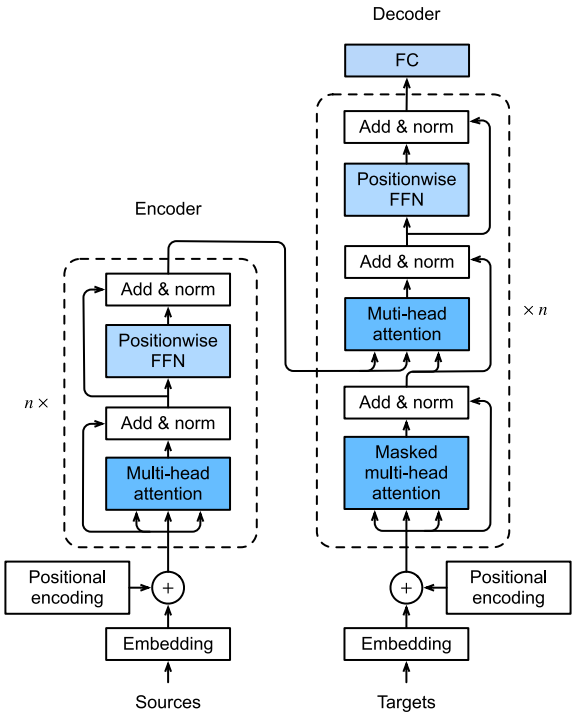


Figure 4: The transformer network architecture.

4.2.2 Text-expression Translation Module

To enable real-time facial expressions for Sophia while chatting with humans, we utilize a methodology inspired by the FaceFormer model[WZZ22]. FaceFormer is a state-of-the-art approach for speech-driven 3D facial animation using a Transformer-based autoregressive model. It processes raw audio input and generates a sequence of animated 3D face meshes by encoding long-term audio context and autoregressively predicting facial movements. Key features of FaceFormer include the use of self-supervised pre-trained speech models for robust audio feature extraction and specialized attention mechanisms to align audio and facial motion modalities effectively.

Our pipeline builds on these principles to generate facial expressions from text inputs. Here's how our pipeline works:

- Text to 3D Mesh Conversion. We employ a text-to-3D mesh model similar to FaceFormer, where the input is text rather than audio. The text input is processed using a pre-trained language model, such as GPT-4, to generate intermediate embeddings that capture the semantic content of the text. By post-processing, we can get ARKit parameters from 3D face meshes.

- Real-Time Expression Generation. With the ARKit blendshape parameters, Sophia's facial motor system can generate real-time expressions. This process is synchronized with the conversational responses generated by ChatGPT, providing a seamless and interactive experience for users. As humans chat with Sophia, they can see her facial expressions change in real-time, reflecting the emotional tone and content of the conversation.

This pipeline leverages advanced deep learning techniques and the principles of the FaceFormer model to enhance Sophia's interactive capabilities. By bridging the gap between textual input and dynamic facial expressions, we aim to create a more engaging and lifelike interaction experience with humanoid robots. The results are shown in 6.

4.3 ROS package and code explaination

For the project aimed at enhancing the expression control of Sophia Robot, the **openai_ros** package presents a powerful platform for training and evaluating deep reinforcement learning or self-supervised learning models. Utilizing this package, we can simulate Sophia's facial expression mechanisms and the associated control system within a ROS environment, allowing for extensive experimentation and optimization. Furthermore, the ability of **openai_ros** to integrate with real-world data and feedback loops enables the trained models to be finely tuned for accuracy and responsiveness.

For the mapping network from blendshape parameters **b** and motor parameters **m**. We use the PyTorch framework to do the training optimization and evaluation. The network is Transformer with embedding size 64,and the optimizer is Adam. We set the learning rate at 1e-5. The total learning epoch is 1000. For our recorded expression dataset, there are around 108k pair-wise data. In the end, there are around 75.6k frames for Stage I's training. We use ROS's recording function to record the values of Sophia's motor parameters and their corresponding timestamps.

5 System Evaluation

5.1 Expression Accuracy

Experiment. We trained our model using a dataset of approximately 30,000 expressions, each with corresponding ground truth motor parameters and ARKit parameters. The dataset was split into an 80% training set and a 20% test set. For evaluation, we recorded Sophia performing a set of predefined expressions and captured the corresponding ARKit parameters. We compared these parameters with the target ARKit parameters using mean absolute error (MAE) as the similarity metric.

Key Performance Measure. The primary measure of performance was the similarity score between Sophia's expressions and the target ARKit parameters, quantified by the MAE.

Success Criteria. Success was defined as achieving a low loss (e.g., >0.9) between Sophia's expressions and the ground truth ARKit expressions, indicating accurate reproduction of facial movements. We can see that, the mae loss in the table and figure is relative small to the motor params' range from (-1, 1). This indicates that our training is a success.

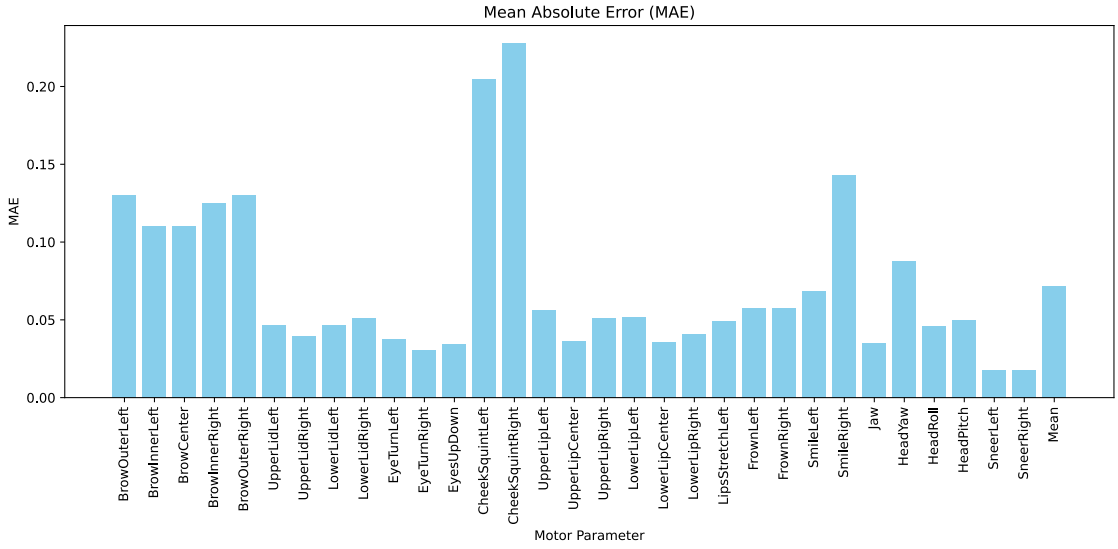


Figure 5: Comparison of predicted and ground truth motor parameters using MAE.

The results are summarized in the figure 5. The table presents the calculated loss values, and the figure illustrates the loss across all 30 motor parameters.

(a) Directly Control of Sophia's Facial Expression with iPhone, and Apple ARKit.

(b) Sophia Imitates Actors' performance from film clips.

(c) Chat with Sophia, using GPT-4o and Azure TTS Service.

Figure 6: Various Interactions with Sophia.

Analysis: Upon reviewing the loss values, we observed that the two checksquint motors had relatively higher losses compared to other motors. Additionally, the smileright motor exhibited a notably high loss. The elevated loss for the checksquint motors could be attributed to the complexity and subtlety of these expressions, which might be more challenging for the model to replicate accurately. In the case of the smileright motor, the high loss is likely due to a mechanical issue, as the motor is currently broken and needs to be fixed. This discrepancy highlights the importance of maintaining the mechanical integrity of the robot to ensure accurate expression replication.

5.2 Expression Range

Experiment. Provide Sophia with a series of commands to express a variety of emotions spanning a broad spectrum (e.g., happiness, sadness, surprise) and record her responses.

Key Performance Measure. Number of distinct expressions successfully performed by Sophia.

Success Criteria. Sophia demonstrates the ability to produce a wide range of emotions, covering the intended emotional spectrum.

In Fig.2, we show our preliminary captured dataset. It contains a diversity of expressions, which can be useful for our training process.

5.3 Naturalness

Experiment. We evaluated the naturalness of Sophia's expressions by driving her with a real human, a film clip and through real-time conversations. Videos of these expressions were shown

to human observers, who were then asked to rate the naturalness of each expression on a scale from 1 to 10.

Key Performance Measure. The key performance measure was the average rating of naturalness for each expression, collected from the observers.

Success Criteria. Success was defined as achieving a high average rating across all expressions, indicating that observers perceive Sophia's expressions as natural and lifelike.

Expression	Film Clip Rating (Avg)	Real-Time Chat Rating (Avg)
Smile	9.2	8.8
Surprise	8.7	8.5
Anger	8.5	8.3
Sadness	9.0	8.6
Neutral	8.8	8.4
Overall Average	**8.84**	**8.52**

Table 1: Average ratings of naturalness for Sophia's expressions as evaluated by human observers.

Analysis: The results from the rating table indicate that Sophia's expressions were perceived as highly natural by human observers, with an overall average rating of 8.84 for the film clip-driven expressions and 8.52 for the real-time chat-driven expressions. These high ratings suggest that our approach successfully enhances the naturalness and lifelike quality of Sophia's facial expressions. The slight variation between film clip and real-time chat ratings could be due to the more controlled and polished nature of pre-recorded film clips compared to the spontaneous nature of real-time interactions.

5.4 Application and Results

We present three key demonstrations as results, shown in Fig. 6. Firstly, we utilize iPhone ARKit to control Sophia's facial expressions in real-time. ARKit tracks facial movements via the phone's camera, mapping these parameters to Sophia's motor system for dynamic replication of user expressions. It significantly improves the naturalness of human-robot interactions by providing a responsive and empathetic interface. Secondly, we enable Sophia to imitate actors' performances from movie clips using a transformer-based architecture. This system accurately maps facial expressions from ARKit parameters from video footages to Sophia's motor parameters, allowing for precise reproduction of subtle emotions depicted by the actors. This capability highlights the potential for nuanced emotional expressions in enhancing human-robot interaction. Lastly, we enhance Sophia's conversational abilities using the latest GPT-4o language model and Microsoft Azure's Text-to-Speech (TTS) API. This integration allows Sophia to engage in dynamic conversations, with real-time facial expressions reflecting the emotional context of the dialogue. These advancements collectively contribute to creating more lifelike and interactive humanoid robots, with potential applications in customer service, entertainment, and personal companionship.

6 How To

6.1 Hardware Information

The Sophia Robot used in this project is located at the MARS Lab, School of Creativity and Arts. Sophia is a humanoid robot equipped with an embedded microcomputer with ubuntu system that controls its various functions, including facial expression mimicry.

6.2 Software and Code

All the necessary code and scripts for running the facial expression control system are stored on the embedded microcomputer within Sophia. No external hardware or additional software installations are required as the environment is pre-configured.

6.3 Step-by-Step Instructions

6.3.1 Powering On Sophia Robot

1. **Turn on Sophia Robot's Power**:
   - Locate the main power switch on Sophia's power system.
   - Turn on the power switch.
   - Wait for approximately 30 seconds to ensure the system initializes properly.
2. **Activate the Sophia System**:
   - Locate the switch for the Sophia System on Sophia's back.
   - Turn on this switch.
   - Allow the system to boot up, which takes about 60 seconds.

6.3.2 Accessing the Embedded Microcomputer

3. **SSH into Sophia' s Computer**:
   - Use an SSH client to connect to Sophia' s embedded microcomputer.

     ```
     ssh user@sophia_ip_address
     ```

   - Replace `user` with the actual username and `sophia_ip_address` with the IP address assigned to Sophia' s microcomputer. We will fill in this part when we finalize all the things in the final week.
4. **Navigate to the Project Directory**:
   - Once connected via SSH, change the directory to the project folder:

     ```
     cd ~/sophia_dev
     ```

6.3.3 Running the Code

5. **Execute the Facial Expression Demo**:
   - Run the pre-configured Python script to start the facial expression control demo:

     ```
     python demo_llf.py
     ```

6.4 Notes

- **Pre-configured Environment**: The required software environment, including necessary libraries and dependencies, is already set up on Sophia's microcomputer. There are no additional prerequisites or installations needed.
- **Code Repository**: All relevant code is stored locally on Sophia' s embedded microcomputer under the `sophia_dev` directory under the home directory.

By following these steps, you can reproduce the facial expression control project on the Sophia Robot, leveraging its advanced hardware and pre-configured software environment to achieve dynamic and real-time expression replication. We also note that this is only a intermediate version of our Sophia project. The final version is subject to change.

7 Conclusion

This project aimed to enhance the facial expression capabilities of the Sophia Robot, developed by Hanson Robotics, by addressing the limitations of pre-defined, artist-created configurations. Despite Sophia's sophisticated mechanical design, there was a significant gap in the control systems needed for dynamic, real-time expression replication.

We used a dataset of approximately 30,000 expressions with ground truth motor parameters and ARKit parameters to train a Transformer network. This network learned the mapping from blendshape parameters to motor parameters. Inspired by the FaceFormer model, we created a pipeline that translates text input into corresponding ARKit blendshape parameters.

Our methodology demonstrated significant improvements in the fidelity and responsiveness of Sophia's facial expressions. The results showed high accuracy in replicating natural interactions, with detailed analysis revealing some areas for future improvement, such as the checksquint motors and addressing mechanical issues with the smileright motor.

References

[BBM+98] Rodney A Brooks, Cynthia Breazeal, Matthew Marjanović, Brian Scassellati, and Matthew M Williamson. The cog project: Building a humanoid robot. In *International Workshop on Computation for Metaphors, Analogy, and Agents*, pages 52–87. Springer, 1998.

[CB02] Erika Chuang and Chris Bregler. Performance driven facial animation using blendshape interpolation. *Computer Science Technical Report, Stanford University*, 2(2):3, 2002.

[CHL+21] Boyuan Chen, Yuhang Hu, Lianfeng Li, Sara Cummings, and Hod Lipson. Smile like you mean it: Driving animatronic robotic face with learned models. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2739–2746, 2021.

[OHK+06] Jun-Ho Oh, David Hanson, Won-Sup Kim, Young Han, Jung-Yup Kim, and Ill-Woo Park. Design of android type humanoid robot albert hubo. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1428–1433. IEEE, 2006.

[RH16] Fuji Ren and Zhong Huang. Automatic facial expression learning method based on humanoid robot xin-ren. *IEEE Transactions on Human-Machine Systems*, 46(6):810–821, 2016.

[SWJD23] Yi Shi, Jingbo Wang, Xuekun Jiang, and Bo Dai. Controllable motion diffusion model, 2023.

[THA+23] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20621–20631, 2023.

[TZS+16] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[WSKZ18] Olivia Wiles, A Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018.

[WZZ22] Kehan Wang, Jia Zheng, and Zihan Zhou. Neural face identification in a 2d wireframe projection of a manifold object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1622–1631, 2022.

[XLMM+21] Fei Xia, Chengshu Li, Roberto Martín-Martín, Or Litany, Alexander Toshev, and Silvio Savarese. Relmogen: Integrating motion generation in reinforcement learning for mobile manipulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4583–4590. IEEE, 2021.

[ZXCC23] Shaobo Zhang, Qinxiang Xia, Mingxing Chen, and Sizhu Cheng. Multi-objective optimal trajectory planning for robotic arms using deep reinforcement learning. *Sensors*, 23(13):5974, 2023.

[ZZ22] Peng Zhang and Junxia Zhang. Motion generation for walking exoskeleton robot using multiple dynamic movement primitives sequences combined with reinforcement learning. *Robotica*, 40(8):2732–2747, 2022.