iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62019

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Svara Rachana - Audio Driven Facial Expression Synthesis

Karan Khandelwal[1], Krishiv Pandita[2], Kshitij Priyankar[3], Kumar Parakram[4], Tejaswini K[5]
Department of C.S.E, Dr Ambedkar Institute of Technology

Abstract: ***Svara Rachana is a fusion of artificial intelligence and facial animation which aims to revolutionize the field of digital communication. Harnessing the ever-evolving power of neural networks in the form of Long Short-Term Memory (LSTM) model, Svara Rachana offers a cutting edge, interactive web application designed to synchronize human speech with realistic 3D facial animation. Users can upload or record an audio file and upload it to the web interface containing human speech, with the core functionality being the generation of synchronized lip movements on a 3D avatar. The system gives special emphasis on the accuracy of the system to generate reliable facial animation movements. By providing an interactive, human like 3D model, Svara Rachana aims to make machine to human interaction a more impactful experience by blurring the lines between humans and machines.***

Keywords: Facial animation, web application, 3D modelling

I. INTRODUCTION

In the last few years, technology has seen an exponential growth, unpredictable and wild. Before the world knew it, having an online presence had become the mandate, be it for social purposes or professional purposes. Nowadays, any task can have an online equivalent, be it office meetings, interviews or even school classes. Moreover, the rapid development in the field of AI has led to the creation of numerous personal assistants which can help us without day-to-day activities. However, the 2-dimensional interface of these applications lack the distinct feeling associated with authentic human interaction. Many people are wary or even outright uncomfortable with personal assistants and the same goes to online meetings whenever there is an unavailability of cameras. The primary objective of Svara Rachana is to bridge the gap between spoken words and lifelike digital expressions by providing a lifelike 3D model which can be utilised as a person's avatar in an application which authentically mirror spoken words. Svara Rachana utilizes ThreeJs to create personalised avatars to substitute the absence of human interaction and collaborating with a highly skilled LSTM model, is able to provide an accurate and reliable speech synthesis platform. Such a system has a plethora of uses in many industries like entertainment industry, virtual communication and broadcasting platforms, online learning platforms and most importantly, the integration with AI to serve as a catalyst for innovation in digital expression and push the boundaries of machine to human interaction.

II. PROBLEM STATEMENT

A. Literature Survey

In the last decade, the world has seen drastic change in technology which changes the way of our working, the way we communicate. Nowadays, you need not to physically be present at a particular location to communicate with others just you need to click few buttons in your phone and its done, but with advancement in technology, people tend to interact with interactive way, phone calls are neither interactive and nor official so we bring Svara Rachna that not only make our communication interactive but will bring a revolution in deaf communication

The paper [1] compares Babylon.js and Three.js, focusing on performance and ease of use. Minimalistic web apps were created and tested for memory usage and frames per second. Results showed similar performance in frames per second, with Babylon.js using 46% more memory. Ease of use was assessed through hour-long observation sessions with developers implementing basic 3D models. Three.js scored slightly higher in ease of use, but neither received high scores. Despite favoring Babylon.js, frustration was common. The study emphasizes the importance of initial learnability, suggesting it could deter users if not quickly achieved.

The paper [2] focuses on WebGL frameworks, which simplify development by abstracting low-level API calls. It specifically delves into Three.js, a widely used framework, covering its background, getting started, and handling fallback to 2D canvas if WebGL isn't supported.

The chapter explores Three.js API calls for creating cameras, objects, and lighting models, and provides equivalent Three.js code for examples using direct WebGL API calls from previous chapters. Additionally, it introduces tQuery, a library combining Three.js with jQuery selectors.

This paper [3] investigates how to enhance audience immersion in 3D animation by going beyond facial expressions to convey emotions. Through empirical research, it highlights the importance of both facial and body movements in depicting emotions realistically. Using Shapiro's 15 controllers for character animation, the study finds that varied controllers are crucial for optimal realism, with facial and gaze controllers playing key roles across emotional states. A proposed preliminary model, based on basic emotions, guides animators in crafting authentic 3D characters by addressing diverse emotional requirements.

This paper [4] presents an end-to-end hierarchical RNN for skeleton-based action recognition. Instead of treating the entire skeleton as input, the human skeleton is divided into five parts and fed into separate subnets. These subnets hierarchically fuse representations as the layer numbers increases. The final representations are then passed through a single-layer perceptron, with the temporally accumulated output determining the final decision. Compared with five other deep RNN architectures and several other methods on three public datasets, the proposed network demonstrates state-of-the-art performance with high computational efficiency.

This paper [5] discusses the continued relevance of the TIMIT dataset despite its age, with over 20,000 references on Google Scholar. It presents a method for creating TIMIT-like datasets with minimal effort and cost for new languages, addressing the scarcity of comparable datasets, especially for less-documented languages. The success story of this method to languages like standard Thai and Mandarin Chinese, with more collections planned is already known. These datasets will be published through the LDC, accompanied by instructions and open-source tools for replicating the method in other languages.

This paper [6] offers a thorough review of Long Short-Term Memory, a type neural network model which is an advance version of RNN has revolutionized machine learning and neurocomputing. LSTM has notably enhanced Google's speech recognition, Google Translate's machine translations, and Amazon's Alexa performance. It effectively tackles the exploding/vanishing gradient problem, distinguishing it from earlier recurrent neural networks. The paper covers LSTM's formulation, training, relevant applications, and provides code resources for a toy example.

B. Objectives

The project aims to implement a machine learning model with supervised learning to translate human speech audio inputs onto the facial movements of a 3D avatar. Utilizing a LSTM architecture, the audio signals will be transcribed into phonetic units and categorized into custom facial expression classes. Additionally, an end-to-end web application will be developed to enable users to upload files, record in real-time, and generate sequences on a cloud-based server, displaying the final animation on a provided 3D character within the browser interface.

C. Existing System

Existing facial animation methods encompass Traditional Keyframe Animation, where manual creation of individual frames represents various expressions. However, no system is a perfect system the limitations in this is in capturing natural fluidity. Morph-Target Animation morphs predefined expressions but lacks real-time responsiveness. Recent advancements integrate AI and machine learning, utilizing CNNs, RNNs, and GANs for highly realistic animations. Commercial software combines traditional and AI-driven methods, providing a comprehensive suite of tools. These advancements aim to boost the quality and realism of facial animations, addressing the limitations of manual and morph-target approaches.

D. Proposed System

In response to the advancements in facial animation systems, our proposed system introduces a forward-thinking approach leveraging cutting-edge technologies to enhance user experiences. Developed using ThreeJS as the frontend framework, our system incorporates a 3D model of a human-like face, elevating realism and expressiveness in digital avatars. Through seamless integration of audio input and LSTM neural networks, our system analyses audio input to generate synchronized lip movements on the 3D model, bridging the gap between spoken words and lifelike digital expressions.

III. EXPERIMENT

For this project, both Jupyter Notebook and Visual Studio Code (VS Code) serve as the integrated development environments (IDEs), offering versatile platforms for code development, experimentation, and documentation.

Jupyter Notebook facilitates interactive coding and visualization, particularly useful for prototyping and exploring data-driven solutions. Meanwhile, Visual Studio Code provides a comprehensive development environment with extensive support for web development tools, machine learning frameworks, and version control integration.

The dataset utilized for training and testing purposes is the TIMIT dataset, consisting of audio recordings and corresponding phonetic annotations. The system configuration requires high-performance hardware with GPU support and software including Python, TensorFlow, and associated libraries.

Model training involves building a Bidirectional LSTM neural network with dropout layers, aiming to classify phonetic units based on audio features. The trained model achieves an accuracy of approximately 73% on phonetic classification tasks. Furthermore, visualization of classification results through confusion matrices showcases the model's performance across various phonetic categories. Finally, the model's utility extends to translating phonetic units into corresponding viseme classifications, providing insights into speech-to-facial expression mapping.

IV. METHODOLOGY

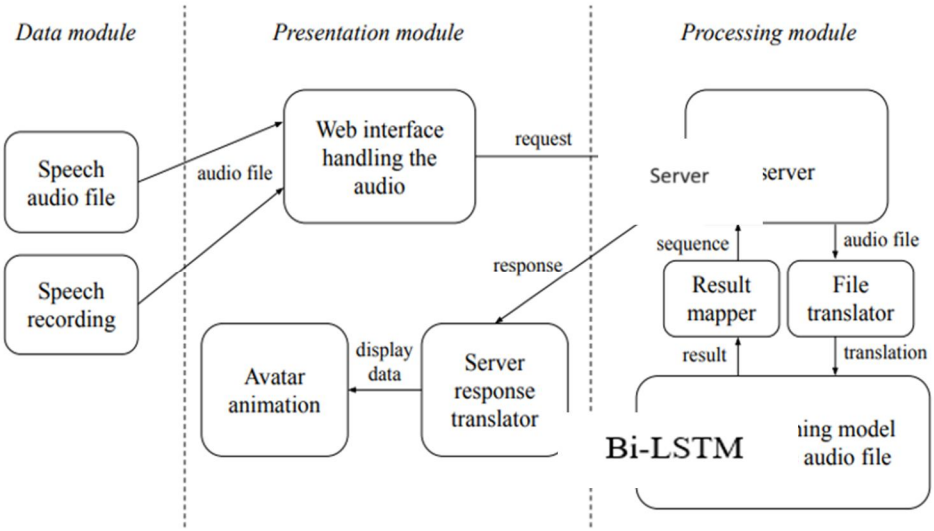


Fig: 3.1

The project is divided into three sections. These sections are built-of of multiple individual nodes which have their own functions.

1) Data module: Users provide audio files or speech.
2) Presentation module: GUI for visualizing the facial animation of an avatar.
3) Processing module: Server, processing the requests using the machine learning model.

The functions of each individual nodes are as follows:

a) Speech audio file and Speech recording: The input is an audio file containing human speech, which the user provides, which can be a pre-recorded file or a new recording in the application itself which the users provide.
b) Web interface handling audio: The UI is utilized for uploading the audio file and recording the voice. If the uploaded file is pre-recorded, it checks if the format of the file is correct or not. Then a request with the audio data is created and sent to the server.
c) Server: The server receives the request from the front-end with the unprocessed audio data. The audio data is then forwarded to the machine learning model through the file translator and then waits for the output. If the file is processed successfully then the mapped results are sent to the presentation module. Otherwise, an appropriate error message is sent.
d) File translator: Here, the audio input is decomposed into relevant coefficients in accordance with the input specifications of the machine learning model.
e) Machine learning model analysing the audio file: The model performs the computations on the received data from the file translator and returns the result.
f) Result mapper: After the computation, the model's result is applied to map the speech phonetics to individual visemes.
g) Server response translator: The server's response is processed and mapped to the coordinates of the 3D avatar.

h) Avatar animation: The translated instructions from the server are applied to the avatar, producing a facial animation.

The machine learning model that is being applied on the project for training the model and testing of the model on the dataset is an advance version of a RNN (Recurrent Neural Networks) that is LSTM (Long Short-Term Memory). LSTM model introduces expressions to the RNN, in particular gates. These gates are of three types:

- Forget gate: This gate regulates the amount of information received by the neuron in the current step from the neuron from the previous step.
- Update (Input) gate: This gate decides whether the neuron will be updated. Also, it controls the amount information the current neuron will receive from a potentially new memory cell.
- Output gate: This gate controls the content of the next hidden state.

The project uses the Bidirectional LSTM for training of the model as well as testing of the model. The difference between BiLSTM (Bidirectional LSTM) and a standard LSTM lies in the input flowing in both direction that is forward and reverse. This model is designed with two layers, one with 256 nodes and other with 512 nodes both of which are a dense layer with the activation function ReLU (Rectified Linear Unit), whose purpose is to extract relevant information that is to be send to BiLSTM layer. The BiLSTM layer has 512 nodes and is the most crucial layer. And the final layer is of 61 nodes which is also a dense layer, one for each phone with softmax activation function. For optimization of the model Adam algorithm is used.

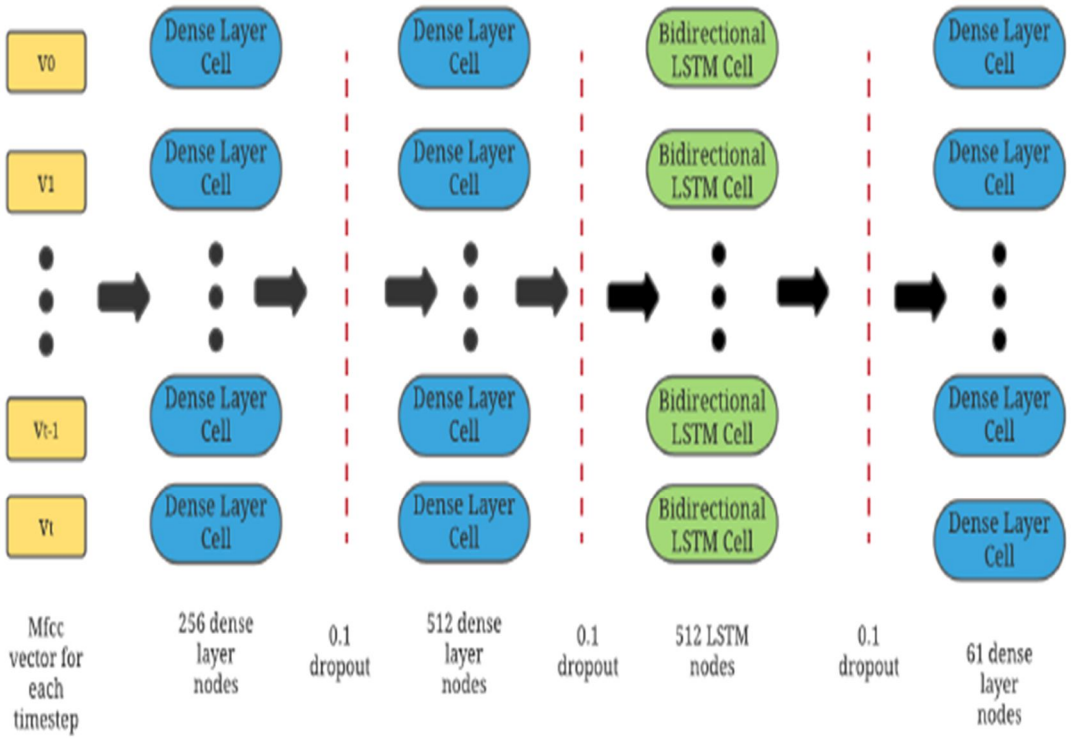


Fig: 3.2 LSTM model architecture

V. RESULTS AND DISCUSSION

The results of our research and experimentation showcase promising advancements in the realm of speech-to-facial-animation synthesis. Through the implementation of a Bidirectional LSTM model, we achieved notable accuracy in classifying phonetic units and mapping them to corresponding visemes. Visualization of classification results demonstrated the model's efficacy across various phonetic categories, laying a solid foundation for realistic facial animation generation. However, challenges persisted in dataset availability and quality, influencing the model's ability to accurately predict facial configurations in some instances. These setbacks underscore the importance of further refining dataset acquisition and processing techniques to enhance model performance. Despite these challenges, our findings highlight the potential of leveraging machine learning in revolutionizing digital expression, paving the way for more lifelike and engaging human-machine interactions.

Fig 5.1

Fig 5.2

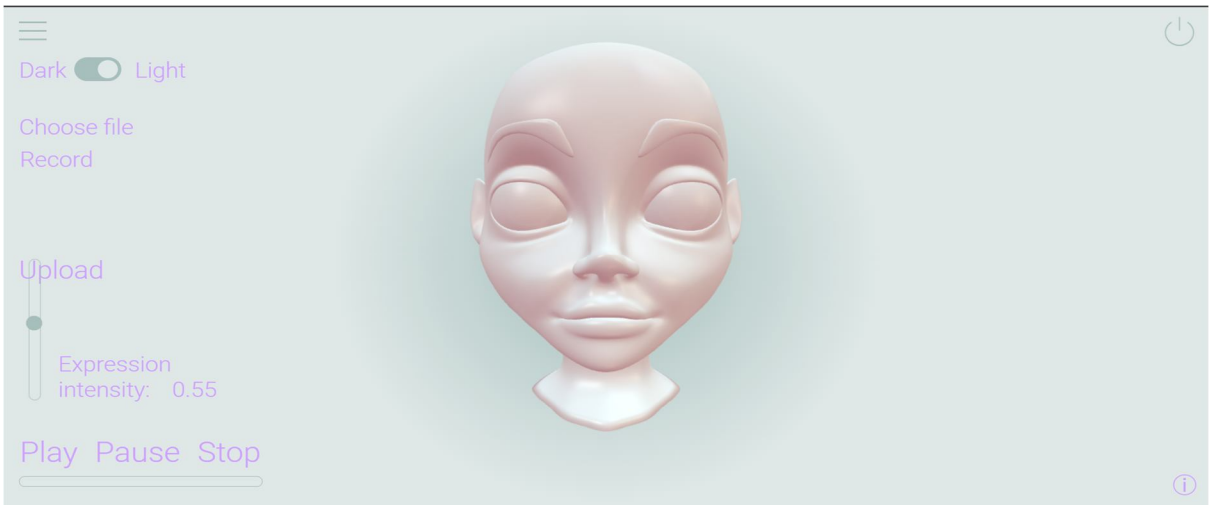


Fig 5.3

Fig 5.4

VI. CONCLUSIONS

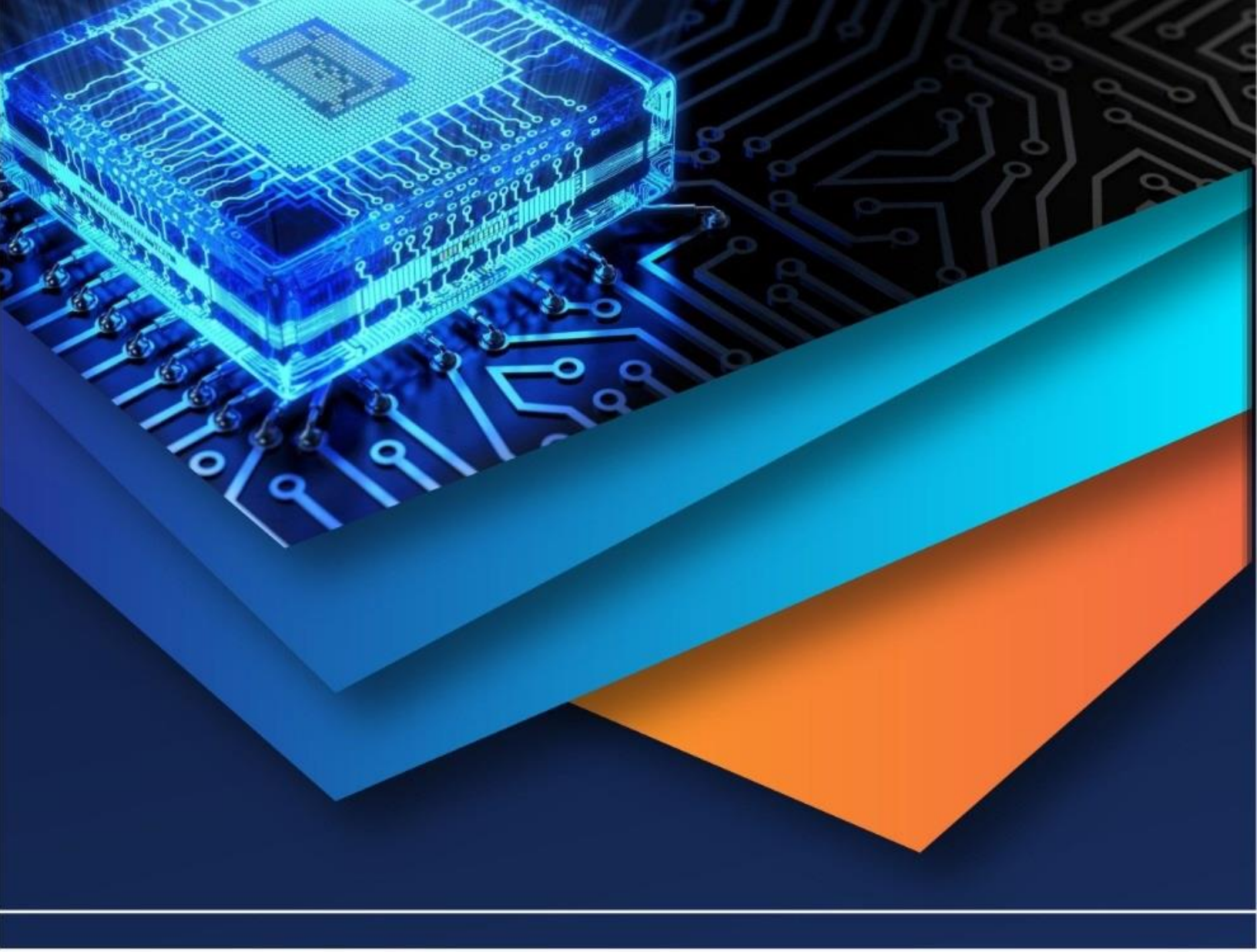
The journey towards implementing machine learning model with supervised learning for speech-to-facial-animation synthesis has been one of discovery and adaptation. Our exploration of the SW-DNN (Sliding Window Deep Neural Network) revealed challenges inherent in dataset availability and quality, hindering the model's ability to accurately predict facial configurations. Despite efforts to mitigate these challenges, including experimentation with different datasets and processing techniques, the model consistently fell short of expectations, predicting mean values for the output space. This setback prompted a critical revaluation of our approach, emphasizing the importance of dataset size and quality in model training. In parallel, the adaptation of animation production techniques underscored the need for nuanced approaches to mapping model outputs to facial bone configurations. Despite initial setbacks, our journey has provided valuable insights into the complexities of machine learning model training and animation production. Moving forward, addressing the challenges encountered, particularly in dataset acquisition and quality, will be essential to refining our approach and unlocking the full potential of speech-to-facial-animation synthesis.

VII. ACKNOWLEDGMENT

We extend our heartfelt appreciation to Prof. Tejaswini K., our esteemed guide and mentor, for his invaluable guidance and unwavering support throughout the duration of this project. His expertise in the field of machine learning and facial animation synthesis has been instrumental in shaping the direction and success of our endeavour. Additionally, we express our sincere gratitude to all those who contributed to the realization of our project, from colleagues who offered insightful feedback to collaborators who provided technical assistance. Special recognition is extended to our project supervisor, whose continuous guidance ensured that our efforts remained focused and productive. Furthermore, we extend immense thanks to our dedicated team members whose diligent work and collaboration facilitated the seamless integration of various components, from developing the machine learning model to implementing the end-to-end web application.

REFERENCES

[1] J. Johansson, “Performance and Ease of Use in 3D on the Web : Comparing Babylon.js with Three.js,” Dissertation, 2021.
[2] Danchilla, Brian. (2012). Three.js Framework. 10.1007/978-1-4302-3997-0_7.
[3] Azlan, Noorsyuhada & Asli, Mohammad & Hamzah, Muzaffar. (2024). Preliminary Emotion-Based Model for Realistic 3D Animation. ITM Web of Conferences. 63. 10.1051/itmconf/20246301021.
[4] Yong Du, W. Wang and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 1110-1118, doi: 10.1109/CVPR.2015.7298714.
[5] Chanchaochai, Nattanun & Cieri, Christopher & Debrah, Japhet & Ding, Hongwei & Jiang, Yue & Liao, Sishi & Liberman, Mark & Wright, Jonathan & Yuan, Jiahong & Zhan, Juhong & Zhan, Yuqing. (2018). GlobalTIMIT: Acoustic-Phonetic Datasets for the World’s Languages. 192-196. 10.21437/Interspeech.2018-1185.
[6] Van Houdt, Greg & Mosquera, Carlos & Nápoles, Gonzalo. (2020). A Review on the Long Short-Term Memory Model. Artificial Intelligence Review. 53. 10.1007/s10462-020-09838-1.
[7] Ruijie Huang, Chenji Wei, Baohua Wang, Jian Yang, Xin Xu, Suwei Wu, Suqi Huang, Well performance prediction based on Long Short-Term Memory (LSTM) neural network, Journal of Petroleum Science and Engineering, Volume 208, Part D, 2022, 109686, ISSN 0920-4105.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY