AN IMPROVED TEXT-TO-GESTURE GENERATION MODEL BASED ON A HYBRID DEEP LEARNING APPROACH

By

TEHMEEMA IRFAN

NATIONAL UNIVERSITY OF MODERN LANGUAGES

ISLAMABAD

August, 2023

An Improved Text to Gesture Generation Model based on a Hybrid Deep Learning Approach

By

TEHMEEMA IRFAN

MCS, National University of Modern Languages, Islamabad, 2019

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In **Computer Science**

To

FACULTY OF ENGINEERING & COMPUTER SCIENCE

NATIONAL UNIVERSITY OF MODERN LANGUAGES ISLAMABAD

THESIS AND DEFENSE APPROVAL FORM

The undersigned certify that they have read the following thesis, examined the defense, are satisfied with overall exam performance, and recommend the thesis to the Faculty of Engineering and Computer Sciences for acceptance.

Thesis Title: A Text to Gesture Generation Model based on a hybrid Deep Learning Approach

Submitted By: Tehmeema Irfan **Registration #:** 62 MS/CS/F21

Master of Science in Computer Science (MSCS)
Degree Name in Full

Computer Science
Name of Discipline

Dr. Moeenuddin Tariq
Research Supervisor _____ Signature of Research Supervisor

Mr. Farhad M. Riaz
Research Co-Supervisor _____ Signature of Research Co-Supervisor

Dr. Sajjad Haider
Head of Department (CS) _____ Signature of HoD (CS)

Dr, M. Noman Malik
Name of Dean (FE&CS) _____ Signature of Dean (FE&CS)

August 16th, 2023

AUTHOR'S DECLARATION

I Tehmeema Irfan

Daughter of Irfan Munawar

Registration # 62 MS/CS/F21

Discipline Computer Science

Candidate of **Master of Science in Computer Science (MSCS)** at the National University of Modern Languages do hereby declare that the thesis **An Improved Text to Gesture Generation Model based on a Hybrid Deep Learning Approach** submitted by me in partial fulfillment of MSCS degree, is my original work, and has not been submitted or published earlier. I also solemnly declare that it shall not, in future, be submitted by me for obtaining any other degree from this or any other university or institution. I also understand that if evidence of plagiarism is found in my thesis/dissertation at any stage, even after the award of a degree, the work may be cancelled and the degree revoked.

Signature of Candidate

Tehmeema Irfan

Name of Candidate

16th August, 2023

Date

ABSTRACT

Title: An improved Text to Gesture generation model Based on a Hybrid Deep Learning Approach.

Non-verbal cues play a pivotal role in providing human-like interaction with Artificial Intelligent Machines like Robots and Digital Assistants. The goal to give machines a human-like aptness is under research for years. Previous work to generate gestures is mostly from speech and also those models are subject to limitations like speaker dependency and gesture quality. This research is to target such problems and develop a gesture-generation model that is capable of producing quality gestures against a sequence of text input words independent of any speaker. Altering an existing specific speaker dataset, a new dataset is prepared including words and gestures. Integration of a sequential Long Short Term Memory algorithm in the model has improved the accuracy of the gestures in terms of Percentage of Corrected Key Points. The experiment results and comparison with relevant schemes show the achievement of the proposed model as it has maximized the PCK and minimized the error rate.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CNN	-	Convolutional Neural Network
LSTM	-	Long-Short term Memory Networks
PCK	-	Percentage of Corrected Key points
MAE	-	Mean Absolute Error
MSE	-	Mean Squared Error
BLEU	-	Bilingual Evaluation Understudy
ROUGE	-	Recall-Oriented Understudy for Gisting Evaluation
ASL	-	American Sign Language
BSL	-	British Sign Language
SSIM	-	Structural Similarity Index

LIST OF SYMBOLS

α	-	PCK Threshold
σ	-	Sigmoid Activation Function

ACKNOWLEDGEMENT

First and foremost, I want to thank Almighty Allah for granting me knowledge, fortitude, good health, and clarity of mind. He enabled me to finish the research.

Without the support and encouragement given from many sources, for which I am sincerely grateful, this research thesis would not have been completed. My research supervisor Assistant Professor Dr. Moeenuddin Tariq and co-supervisor Mr. Farhad M. Riaz, who never stopped guiding me along the way, were one of many steadfast supporters and important contributions to my accomplishment. I want to express my gratitude to my supervisor in particular for his help, direction, and insightful comments on the research.

A special and heartiest gratitude to my worthy teacher Mr. Mohibullah Khan for providing a platform to accomplish and implement this research and guiding me with the best to overcome my obstacles.

Last but not least, I won't forget the management of the Department of Computer Sciences' generous assistance in helping me to complete this research. Many thanks to everyone whom I didn't include but who helped me stay motivated and made efforts I won't forget.

DEDICATION

I dedicate this thesis to my parents, teachers, and siblings, whose love, care, encouragement, and leadership over the years established the groundwork for me to successfully accomplish any task with diligence. I also dedicate this work to my instructors and mentors who provided me with valuable lessons and seek to be the best.

CHAPTER 1

INTRODUCTION

1.1 Overview

Non-verbal cues give meaning to any individual's words, they include movements of hands, gazing, facial expressions, nodding, gestures, etc. Specifically, the hand gestures used while uttering words promote the understandability of content in human-human communication. This phenomenon plays a significant role in human-machine interaction as well. Humanoid Robots have an appearance like human beings and are expected to communicate in the way humans do. Similarly, virtual agents have a human-like appearance and use hand gestures while dealing or interacting with humans. Deep learning has been a great platform for generating these Robotic gestures. As for humans, the central part Brain acts as an intermediate between speech and gestures and instructs which gesture should be generated for a specific word. For machines that central part is Deep Learning which acts as an intermediate between uttered words and gestures for them. Deep learning is a branch of Artificial Intelligence that percepts the human brain and gives the ability to behave like humans to machines. Many approaches have been made to develop a model which generates gestures. Typical rule-based methods are also there which were based on heuristic rules but these have been successful in limited situations. Deep Learning based Data-driven approaches have overcome those limitations. Deep Learning models have played a major role in generating 2D joint coordinates of human gestures. The motivation for this research is to develop a Hybrid Text-to-Gesture generation model using a sequential model to overcome the problems in existing models.

1.2 Motivation

The motivation behind this research stems from the growing importance of gesture recognition in human-computer interaction and the need for more accurate and efficient systems. As technology continues to integrate with our daily lives, the ability to accurately interpret human gestures is crucial for applications ranging from accessibility and communication for people with disabilities to immersive virtual reality experiences. Artificial Intelligence Especially Neural Networks has proven a core in the Intelligent Control of Industrial process. These various Intelligent tutoring systems in education are being used widely almost all over the world. Intelligent Robots are extensively being used for health care, restaurants, and in homes as well. In short, we can say it has made life easy and automated. Since the use of automated machines is becoming common the need to get interact with such machines is a big challenge. One way to tackle this daring task is the need for proficiency in this field for each and every single consumer of technology but its disadvantages are more than advantages. The need to put effort may lead people to revolt. However, the Graphical User Interface is a key to solve the problem. Automated machines should have an Interactive and user-friendly interface to remove all ambiguities to be used in an easy way. Though this phenomenon was proven in increasing human-machine interaction, this discussion is still under the concentration of researchers. The way of human-human communication such as the use of body parts like hands, arms, and facial expressions while communicating and talking with each other has been proven to better understanding of speech. These movements of the upper body parts are known as gestures. Using these gestures can surely increase the naturalness of human-machine interactions. People can communicate and interact with automated machines in an intrinsic way. Automated machines such as Virtual Agents and Robots especially humanoid robots have human-like appearances and they are expected to behave as humans do.

The researchers concerned with the human-computer interface are focusing to involve nonverbal communicative modalities in Human-Machine Interaction. When non-verbal cues are added with a corresponding speech in computer-generated animation or mapping of a humanoid robot they are contemplated more lifelike by humans. This enhances the attributes of social communication in artificial agents. [1] The ingredients of communicative skills such as movements of Hands, arms, and facial expressions are best candidates in increasing the abilities of such agents with respect to the communication. These non-verbal cues and sign language are

part of a continuously moving cycle and can be taken as two ends of continuity [2] and the middle space is filled with two frontiers “Italianate” and pantomime in the company of vocabulary. The spoken words typically called speech along with gestures can foster the understanding of the content and add meanings to the keywords in speech [3] and a positive outcome has come out in human Interactions with virtual agents [4-7]. This practice should be made when ding mapping of a humanoid robot in order to give naturalness to them. As robots particularly Humanoid robots have appearances as humans and they are expected to communicate and behave in a way as humans do. By saying communicating like humans they are expected to use their hand and arms and express their facial expressions while interacting with humans. This spontaneous and corresponding movements of hands and speech can be modeled to promote Artificial Intelligence. [8]. The desire to enhance the PCK of gestures and reduce MAE is driven by the potential to create more intuitive and seamless interfaces, making technology more accessible and user-friendly for a broader range of individuals. This research aims to contribute to these goals by addressing key challenges in artificial gesture recognition, ultimately improving the user experience and expanding the utility of gesture-based technology.

1.3 Robotic Gesture

A Gesture is the movement of upper body parts such as hands, arms and Shoulders doing impulsively while speaking. The mapping of these movements into 2D Key points continuously refers to as an artificial gesture

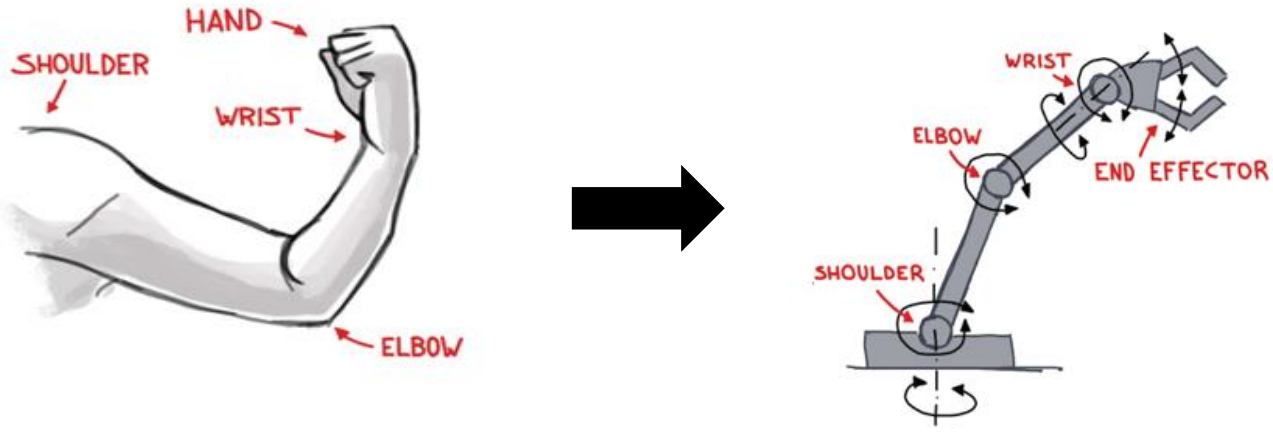

Figure 1.1 An artificial and real parts of an arm. [8]

An artificial arm is comprised of several parts such as shoulders, elbow, wrist and end effectors which perform the role of hand. A sample of joint coordinates of human body is presented in Fig 1.2.

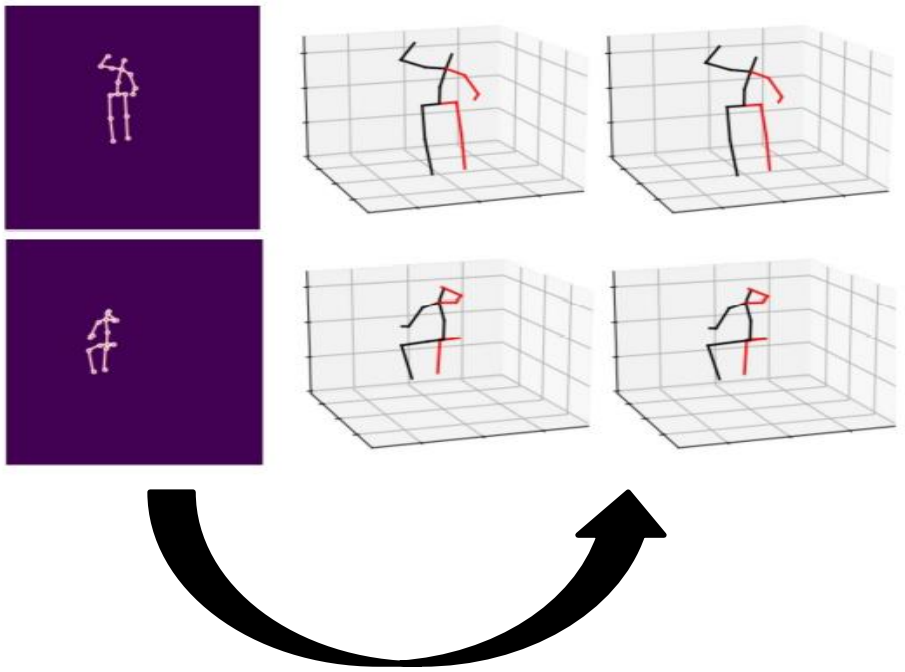

Figure 1.2 Joint coordinates [9]

There are several methods available for mapping real hands into artificial hands. An efficient method to control an artificial robot hand by EEG Signal is presented in [10]. Artificial gestures were produced via myoelectric signal which was obtained by putting electrodes on the muscular skin. The scope was limited to six gestures corresponding to finger's movement, spreading and grasping. A three layered Artificial Neural Network was used for further classification tasks. The produced gestures were executed on an artificial hand of a robot. Recently to estimate human pose in three dimensions Multi-Layer perceptron was used in [9]. The presented pose estimation was mapped from 2D-3D using three modules such as PEM, gJCM, and, PLM. The joint coordinates were embedded in the first PEM (Pose Estimation Module) in the form of $N \times 2$ matrix instead of presenting all the joint coordinates in a single merged vector as done in [11]. Because of this embedding the relation among all the joints were easily modeled. Further the second module gJCM (Joint-Coordinate Mixing with Gating) which is the core of the differences of the presented work, composed of further two sub modules Self-Gated joint-mixing and Selective gated coordinate-mixing and last PLM Pose lifting module

finally mapped the 2D joints into 3 Dimensions as demonstrated in Fig 1.2. The successful implementation of the model was performed on datasets Human3.6M [12], HumanEva-I [13] and, U3DPW [14].

1.3.1 Applications of Robotic/Artificial Gestures

Artificial generated Gestures are view as promising solution in such areas where human approach is difficult and to perform several complex tasks which require more time and energy if done by human themselves. Also the intelligent machines in the form of hands embedded programmed gestures are proven as easy way to interact with systems because of their easy and friendly interface. Some of such applications are listed below:

i. *Medical Systems and assistive technologies:* The distribution of resources in hospitals specifically the interaction with medical instruments can be controlled by Artificial gestures. Gestures can be the part of reconstruction therapy of people having disability like handicapped users.[15] A staying alive Virtual-reality-imagery-and –relaxation tool allows patients of the dangerous disease cancer using 18 traditional T'ai Chi gestures to traverse through a Virtual scene.[16]

ii. *FAce MOUSe:* [15] presented a novel human-machine interface which the position of a laparoscope during laparoscopy which helped surgeons to check and maintain the position of the laparoscope with facial gestures.

iii. *Gestix:* In surgical operations surgeons use this hand gesture tracking device for browsing MRI images.[17] This device provides a natural interface presented in Fig. 1.3

iv. *WearIT@work:* The "comfort" condition is met by the WearIT@work European Community Project, which encourages doctors to utilize a wrist-mounted RFID reader to recognize the patient and communicate with the hospital information system using gestures to record exams and create prescriptions, assisting in maintaining sterility.[18]

Figure 1.3 Surgeon using Gestix to browse medical images

v. *The gesture Pendant:* The Gesture Pendant is a System based on Gesture Recognition System used in homes to control devices and in medical diagnostic tools it gives supplementary performance.[19]

vi. *Tele rehabilitation System:* This system utilizes gesture's force-feedback of patient in Kinesthetic for treating patients with arm-motion coordination disorders.[20, 21]

vii. *Entertainment:* Most of the existing computer systems are having interactive interfaces which give them a congenial nature and cutting-edge interface such as computer games provide economically rewarding arena. Since users specially teenagers are plunged in any inspiring game-like environment they are eager enough to try new interfaces. The reason behind is that control of such systems is under fingertips accompanying specific gestures.[22] Requiring fast response computer games are based on vision system, the system quickly responds to gestures produced by users because in such systems Computer-Vision algorithms are programmed and embedded which are efficient and Strong enough to recognize any specific produced gesture by user. Therefore, more effort should be made to develop gesture recognition systems with processing image at high rates [23]. Users interact with combatant using their hand gestures in fighting game based on augmented reality. [22] The gesture spot problem was undertaken and solved through voice recognition technique.

viii. *Crises management and disaster Relief:* Gesture controlled systems help in preventing from disasters including natural and human-caused tragedies. Such systems include command-and-control systems which require quick response from users. In designing phase of such systems an emergency response must be placed which give access to a large amount of data accompying any typical interface between the computer and the user. Apart from these applications other applications include multi-user hand gesture interfaces provide entrainment to more than one user at a time. Such algorithms incorporate multi-touch control technology. [18, 24] The developers of such systems focus more on performance of collaborative users. These systems entertain more than one user at one time by giving them a variety of vocabulary of gestures such as manipulating an image include zooming, editing etc.

ix. *The Command Post of the Future:* Pen Based gestures are in this system which provide an adequate act by considering the requirements of "Space of Interaction", "Size of Lexicon", "Fast Learning", "aptitude", and "Number of Hands" involving natural interaction to provide a control in any disaster.[25]

x. *Human-Robot Interaction:* Gestures play a crucial role in providing a friendly interface between user and system. Recognition of gesture is critical part. These artificial gestures involve several geometrical properties to perform navigation for a robot. The gestures are triggered on utterance of a specific keyword. For Example: A keyword "Go-There" correspond to a pointing gesture. Also these gestures are used to manipulate operations. A system is presented in [26] trained on six gestures features. The system triggered an action on recognition of a gesture. Two nodding gestures were identified by combining eight different facial expressions and hand poses in [27]. In [28] to control a hybrid service Robot System HARO-1 a novel approach is demonstrated which satisfies user adaptability and feedback using eight different and unchanged gestures.

1.4 Problem Background

Gesture recognition and Generations is itself a challenging task as it's requisites lead designer to be very conscious regarding specific factors such as Gesture Intuitiveness and spotting. The developed Systems based on Gesture Interface are tested upon multiple measures including Multi-headed systems, Interaction, System Configurability and specially to meet "come as you are" challenge. Gesture Based Interfaces are no doubt performing a core part in decreasing human-machine distances. Yet, they undergo certain challenges. Some of them are listed below:

i. *Costs/Benefits:* Most of the devices having gestures interfaces require valuable cameras and high-power consumption, advanced sensors and its setups and also its acceptance in market is a question. Therefore, considering stable budget systems are needed to be developed with such hardware without which system cannot fulfill requirements.

ii. *Responsiveness:* Real-Time Gesture Recognition with "no-delay" is demand to improve responsiveness in systems otherwise it will not be acceptable system. In [29] maximum latency rate between "System response" and "Event Occurrence" was 45ms. A system producing response after 300ms is considered as inactive.

iii. *User adaptability:* Existing Systems are tuned to their pre-defined algorithms which recognize only designer's selected gestures because of their offline training. Therefore, the generalization is required which can be achieved if online training is given to the system and new samples are added on requirement.

iv. *Learnability:* For controlling applications gesture patterns are required to be easy. To get this achievement the system should adopt natural and intuitive patterns [30] focused on acceleration-based gesture patterns.

v. *Accuracy (detection, tracking, and recognition):* In Detection, Tracking and Recognition Systems accuracy is conditioned. Existing Systems mostly consider

complex measures for assessing. [31] Addressing factors like attributes of hand such as hand skin type and color etc can improve performance.

The Hybrid Deep Learning techniques are well proven in different applications like image and text classification and CNN have shown good performance in text processing and semantic analysis. Therefore, Hybrid Deep learning techniques can be used to generate gestures. The main motivation of this research is to propose the hybrid deep learning gesture model and analyze how a sequential model and text input modality can improve the quality of gestures. Gestures Interface play a crucial role in Human Machine Interactivity. It has been decreasing the constraints on human collaboration to use advanced technologies like Humanoid Robot which are specially designed to perform human activities since recent years. Yet it has certain limitations and it is being undergone with challenges such as the accuracy of produced gestures which defines the quality. [32, 33]. The developed systems are trained to produce a pose against a keyword like for a keyword “Put that There” the gesture should be of pointing. Any kind of other pose, if gets generated by the system will be considered as inaccurate and there exists no chance to consider it. Therefore, for applications like Gesture Recognition and tracking, we cannot compromise on accuracy. The proposed architecture of gestures generation proposed in [32] achieves an accuracy up to 56% in terms of Percentage of corrected key points.

Inspired from human way of communication as they use co-verbal gestures along with speech multiple approaches are there to produce gestures from speech [32, 34, 35] . Apart from speech text is another important entity widely used by Artificial Systems. For Example: A humanoid Robot can be programmed to serve as waiter and deliver desired item from menu by the customer. The customer’s choice can be based on text. Very less literature is found on gesture generation technique from text input modality.[33]

Recently, proposed gestures models are based on speaker specific data [32, 33] either in form of text or audio which consequently limits its scope. Based on the generalization models are so far seen. [36]. Therefore, it is a requirement to design such generic input based models. As produced gestures from any input modality demands a sequence. Deep learning has several sequential models such as Recurrent Neural Networks and Long Short Term Memory which

are used in such applications where long term dependencies are required. Embedding a Sequential Model may lead in increase in quality of output and improve accuracy.

1.5 Problem Statement

While communicating gestures are important part and they convey meaning to speech content. Existing approaches generate gestures for Artificial Intelligent systems like Robot using audio input modality specific to a particular speaker [33, 36]. Along with speech or audio text is a same worth entity for communication specially in Robots which is encountered very rare. This research is to address this problem and propose a gesture model using text input modality and improve quality of gesture via Hybrid Deep Learning approach using a sequential model.

1.6 Research Questions

i. How to modify the Gesture Generation model using Text Input Modality?

ii. How does the Hybrid Deep Learning Sequential (CNN-LSTM) model affect gesture quality?

1.7 Aim of Research

The primary aim of this research work is to advance the field of artificial gesture recognition by improving the PCK and reduce the MAE of generated gestures. Focusing on specific gestures involving arms, hands, wrists, and shoulders and utilizing textual input in the English language, the research seeks to enhance the quality of gesture recognition, ultimately contributing to more effective human-computer interaction and interface design. By addressing the challenges associated with gesture key point prediction and minimizing error rates, this study aims to provide valuable insights and practical solutions for applications such as sign language recognition, virtual reality, and gesture-based control systems.

1.8 Research Objectives

i. To Propose a modified Gesture Generation model using Text Input Modality.

ii. To improve the quality of gestures using Hybrid Deep Learning Sequential (CNN-LSTM) model.

1.9 Scope of Research Work

The scope of the research is limited to specific gestures related to arms, hands, wrists, and shoulders, excluding gestures involving other body parts. Additionally, the recognition of gestures is exclusively based on textual input in the English language, which may not account for the diversity of languages and non-textual input methods commonly used in practical scenarios. The research's effectiveness is contingent on the quality and diversity of the dataset used for training and evaluation, which may present limitations in capturing the full spectrum of real-world gesture variations. Furthermore, hardware and device-specific factors that can influence gesture recognition accuracy are beyond the scope of this research. Despite these limitations, this study aims to contribute valuable insights and improvements to the field of artificial gesture recognition within its defined boundaries.

1.10 Thesis Organization

The rest of the Thesis is organized as follows:

Chapter 2 provides the research domain and detailed overview of the existing literature. Classification of literature with respect to multiple techniques and approaches like models, algorithms and specially methodology used id provided. Further Chapter 2 provides existing constraints and challenges which provide a platform for further research and the research gap which lead to design a modified Text based gesture model.

Chapter 3 grants methodology for this research in a detail, providing details on all the existing benchmark techniques and solutions to cope with those present constraints. This Chapter describes the methodology plan opted for conducting this research. Also it gives details of the implementation tools and how the proposed model will be evaluated.

Chapter 4 can be considered as core of thesis because it provides with in detail view of proposed architecture of model. Discussing about the algorithms, models used and tool to implement the proposed architecture for generating gesture model. Details about data and what necessary data processing requirements are met is discussed in this chapter. Furthermore, the flowcharts, figures and diagrams presented in this chapter for better understanding of model's architecture.

Chapter 5 will provide the reader with evaluation of the proposed model and undertaken parameters to evaluate the put forward architecture of model. Further in this chapter comparison of the proposed model with benchmark datasets is presented in this chapter which guarantees the effectiveness of this architecture. Achieved accuracy using proposed approach against other model is presented in the form of table.

Chapter 6 presents summary of contributions of this research work with directions towards the future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

Lifelike gestures of a humanoid robot have been under consideration by researchers in recent years. In this chapter, a brief review of the state-of-the-art techniques for generating gestures is provided which delivers the obligatory factors and stows a rigid bedrock for the research. Generating Artificial human-like styles is challenging i.e. synchronicity of gestures with voice. Applications of Gesture Recognition are most demanding as they strengthen the flexibility in Human-Machine Interaction. This research effort bestows a distribution and presents a study on Neural Network Based techniques for generating gestures in section 2.2, techniques related to Morphemic Analysis and Generative Adversarial Networks (GANs) are discussed in sections 2.3 and 2.4 respectively. Sections 2.5 and 2.6 contain approaches from Virtual Reality and Data-Driven Approaches respectively to generate gestures. Techniques to generate Signs are discussed in Section 2.7 and a difference between gestures and Signs is provided in Section 2.8. Highlighted Research gaps and directions are discussed in section 2.9. Finally, the chapter is summarized in Section 2.10

2.2 Neural Networks

Artificial Neural Networks tend to process data in a human-energized manner. A Neural Network consists of one Input Layer, an Output Layer, and hidden Layers. An increase in the number of Hidden layers leads to Deep Neural networks.

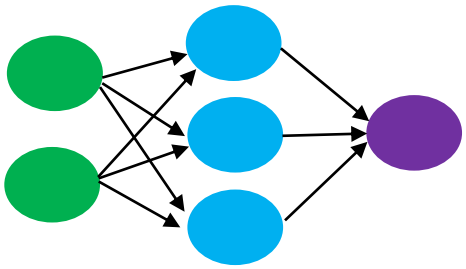

Figure 2.1 Basic Neural Network

Eiichi Asakawa [33] proposed a text-to-gesture generation model and evaluate it with speech-to-gesture using the 1D U-net Neural Network Architecture. The dataset used was based on speaker-specific spoken texts extracted from the videos in [32] using "Cloud Speech to Text" and 49 2D key point coordinates corresponding to the neck, shoulders, wrists, and hands. A special token <BLANK> was assigned in video frames each of 15 fps without any spoken text. The model was designed for the transformation $R^{300*N} \rightarrow R^{98*N}$ [33] where 300 is the number of input feature vectors and 98 is the number of 2D Key points and N represents the frames of the input sequence.

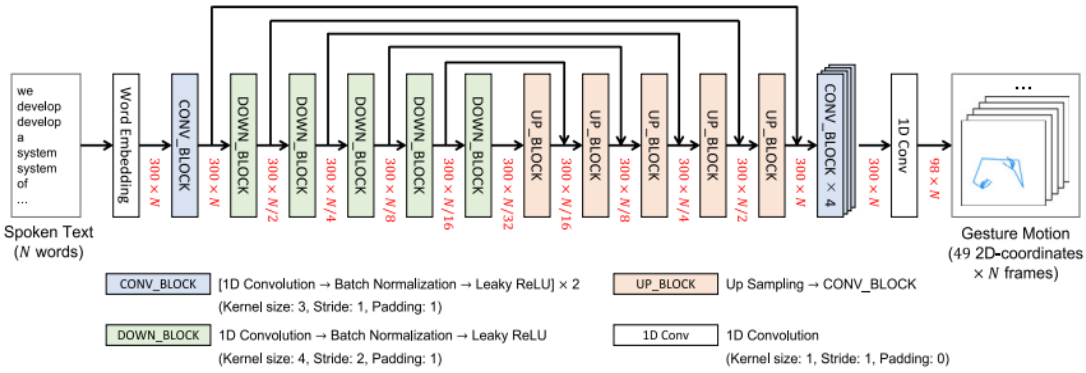

Figure 2.2 Text-to-Gesture model [33]

A skip connection was added to transfer the time series information to the decoder part. The L1 Loss function was used as an optimization function. Ginosar et al [32] suggested a mapping linking audio and pose using a fully connected Artificial Convolutional Neural Network with an audio encoder that compresses the 2D spectrograms of speech and remodels them to a 1D signal. The following U-Net Neural Network Architecture G maps this signal into a temporal stack of pose vectors. A training signal was provided by the Loss Function L_{L1}

subtracting predicted motions from the original ground truth. The real Key-points of ground truth were extracted using Open Pose from the videos of 10 famous gesturing speakers out of which there were 5 talk show hosts, 3 lecturers, and 2 televangelists. The generated motion predictor was compared with Nearest Neighbors and the RNN-based model. Unlike earlier studies that attempt to discover one-to-one mapping of speech-to-gesture, Wu Bowen et al [37] suggested a Conditional Generative Adversarial Network Model for speech-to-gesture conversion and estimate the gesture distribution conditioned on speech via parameterization. The presented work is threefold. First, the co-speech gestures are generated. Secondly, performance was improved by changing the gesture patterns by manipulating the randomly sampled vector. Last, the model's performance is confirmed with the help of a certain objective and subjective experiments. The proposed CGAN architecture takes voice features as the condition. A noise vector (z) is copied from "Gaussian Prior" having the equal length as input Speech features. These features and (z) are further refined by fully connected layers and passed into bi-LSTM. The output is further taken by sequence-wise fully connected layers specifying the absolute positions of each joint in 3D space were generated. The Discriminator was concurrently optimized by calculating the error between the generated and the real distribution.

Chung-Cheng Chiu et al [38] proposed his work of generating gestures using a Data-Driven Approach. Instead of producing gestures of all kinds the model generated motion associated with Prosody based on Hierarchal Factored Conditional Restricted Boltzman Machines (HFCRBMs) [39]. The control of transformation in motion is upon Prosody and learns attributes of human poses and obstructions in an animation via training data consisting of speech and poses humans during conversations. The dynamics of these poses are in detail learned by the model. Finally, the trained model was able to produce animations against a recorded speech.

2.3 Morphemic Analysis

By speech, utterance gestures can be classified into four groups Iconic, Metaphoric, Deictic, and Beat [40] Iconic gestures often portray images of items and volume in steady movement. Metaphoric gestures demonstrate speedy actions. Deictic gestures are denoting gestures and Beat gestures energetically accentuate dominant chunks of an utterance. Yu-Jung Chae et al [41] proposed a technique to generate gestures using this gestures Classification using

morpheme analysis. Morpheme Analysis is distinguishing weighty segments of words from a sentence containing N words. In the proposed work morpheme analysis was performed on the words of the speaker and gestures were planned accordingly to parts of speech and constituents identified using the KKMA morphological analyzer which has an accuracy of up to 75% [42]. Utilizing this outcome, a sentence gets split into one or more expression units where it goes with a phrase or word. The gestures were prognosticated for each expression from Gesture Data Base. Random Forest was used for confirmation of expression units and their gesture types.

Speech and Text simultaneously play an important role in producing gestures. Speech gives us the information about the uttered sentence such as timings and pitch through which we can produce Best gestures [43] which are rapid movement of hands in upward and downward direction and the factor of time can tell us how long should a gesture position e continue whereas from the text we can have semantic meanings of the words. Words can be filtered from a sentence which convey exact meaning or simply main words. Following this a combined approach is proposed in [44] which uses both approaches to produce gestures. Two techniques Part-of –Speech (PoS) based and Prosody-Based were combined to generate appropriate gestures. In first Event Driven Module Sallow Parsing was applied on input text store in SSML (Speech Synthesis Markup Language) [34] for analysis and keywords were identified and emblematic gestures [45] were produced against each keyword hence computed a motion sequence and best gestures were produced against a content word. In second module only beat gestures were produces from the Prosody of speech specifically pitch value for modeling intonation. Two parameters pitch and time points were encountered to represent prosody of beat gesture. Same parameters were extracted from SSML text using mbrola voices [46]. Relevant gestures were produced where these parameters matched. While combining two modules

emblematic gestures were given priority as it has specific meaning and used consciously. An overview of this proposed Architecture can be viewed in Figure 2.3

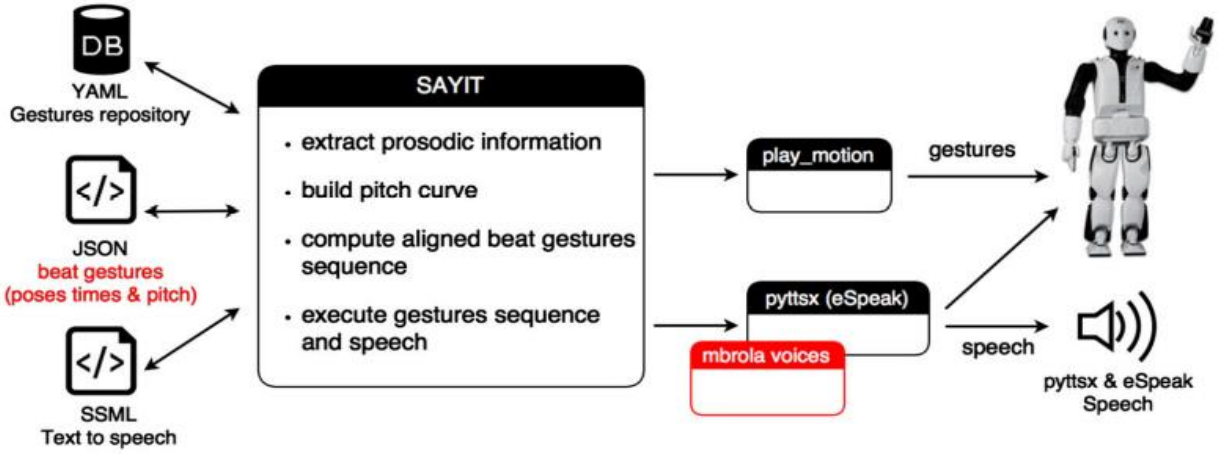

Figure 2.3 An architectural method based on prosody. Each beat gesture in the gestures database has a prosody curve specifically described for it. The gesture with the most similar form is then assigned to each pitch peak of the text to be uttered, and this results in the computation of a motion sequence. No emblematic movements are made.[44]

2.4 Generative Adversarial Networks

Generative Adversarial Networks incorporate two different appended networks one is known as Generator (G) and the second one is Discriminator (D). (G) produces doable data and (D) magistrate this output to discern between real and fake. Unai Zabala et al [47] proposed a methodology for gesture generation using GANs. The second neural network of the model (D) was learned its distribution space with 56 input dimensions for 2000 epochs utilizing 2018 units of movements (UM), each carrying a series of 4 successive poses being in tuned with key point coordinates of the head, hands, wrists, and arms. (G) was sowed with a consistent distribution with a range [48] of 100 dimensions. Tuning of Hyper-parameters was done on probation using [49] as an Optimization strategy. Since speech and gestures are correlated,

gestures were generated via modeling the correlation between speech utterances and poses [50]. The proposed work is multi-model. The 3D poses of gestures were extracted from 2D videos of humans performing actions. While extracting only video fragments were undertaken which entirely present the speaker's arms, hands, and face. The GAN framework generated the gestures by implementing a multi-model task. The target of (G) was to predict gestures similar to ground truth and passed over to another neural network (D). The implementation was based on U-net Architecture presented by [51]. The prosodic audio features were forwarded using skip connections inside the U-net and high-level features containing information about the long sequence of input were extracted via the U-net bottleneck. The model was optimized using the $L1$ loss function. The model generated the sequences of 3D human poses. These produced key points were connected and a skeletal representation of human poses was done in a virtual environment for the animation of an avatar.

While speaking every individual has own way of talking and style of producing gesture spontaneously. For a humanoid robot to learn these individual styles is a challenging task. An attempt has been made to learn individual styles of multiple speaker for a Robot [36]. The proposed model is called Mix-StAGE and learns to produce gestures from an audio input. A single model is learned to produce co-speech gestures of different debaters while preserving the uniqueness of every individual in an end-to-end way. The novelty of the proposed model from other speaker specific gesture generation models [32] [52] [53] [54] is learning and producing styles for general audio input. The task was performed with two jobs first was Style Preservation which made sure that model is still capable to save uniqueness of each style while learning from different speakers and second one is Style Transfer which ensured that produced gestures come from a fresh aesthetic that differs from the speech's original source.

2.5 Virtual Reality

In recent years the approach of visual presentation has outstretched to a high level of puberty. Visually representation of the environment refers to Virtual Reality. This environment is generated with the help of Computer Technology. One can traverse this environment in 360 degrees. To interact with Virtual Environments Marc Erich Latoschik et al [55] proposed his work which aims to produce speech-gesture interfaces to interact with such environments.

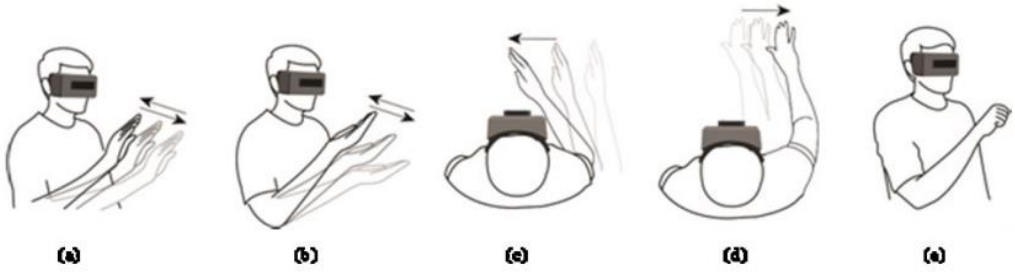

Figure 2.4 All hand gestures for movement control (a) forward (b) backward (c) Left Step (d) Right step (e) Hold Position [56]

Gestures were produced to manipulate objects in a virtual environment projected screen on the wall. This task was accomplished in 2 steps. First for every pointing event to select any object from Virtual Environment Standard Geometric Transformations were applied to map the two coordinate systems. One represents the real world and the other represents the Virtual one. Secondly, to manipulate the selected objects two operations Geometric rotation and Translation were applied to the objects. Further, these gestures were combined with speech to interact with Virtual Environment in a more Robust and Efficient way e.g. combing a pointing gesture with utterance keywords “this” or “that” will narrow down the contenders to select any object.[55]

An effective tool for a superintending string of hand gesture events is the Finite State Machine (FSM). [57],[58] Based on (FSTM) Wanhong Lin et al [59] proposed a novel approach for depicting 3D hand gestures to manipulate items in virtual reality by identifying important features of gesture design and suggested action Determining rules and an action-state transition method. Control of objects in Virtual reality was supported by Leap Motion Hand Sensing Device.

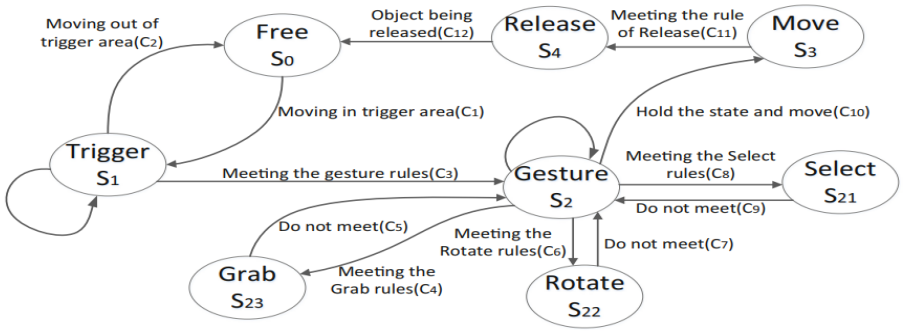

Figure 2.5 The finite State Machine of action. [59]

The proposed finite state action machine is composed of five action states, six input conditions (C) of input event (X) corresponding to states (S). The state transition function $S(t+1) = f(X(t), S(t))$ where t indicates the time, $X(t)\,\epsilon\,X$ and $S(t)\,\epsilon\,S$ [59] Hand gestures features and the distance between the virtual hand and the object should be less than the object action trigger threshold. To satisfy the previous statement defined as non-contact action determined rules were the condition on the object and virtual hand for the successful completion of an action.

2.6 Database-Driven Approach

A Data Driven approach in research refers to extraction of interesting insights in a scientific way from data. This technique requires an extensive amount of data to perform complex calculations. Gestures are created using database from speech [60]. The proposed work is consisted of several strides. Dataset used was comprised of 23000 individual gestures brought out from movement seize data. From the speech input pitch contour speech was examined using Praat [61]. Frequency of a particular gesture was extracted using motion analysis as a gesture with high frequency can be proved fruitful in increasing a Robot's extraversion's perceptions [62]. Following the rule that Peak of speech either succeeds or coexists with gesture peak [63] gesture timings were set in such a way that gesture blows are 55% completed at peak of pitch. Further 5 parameters naming Velocity od Gesture, mean acceleration of the first major velocity peak, measuring size of gesture considering total length of path, swivel of arm nd opening of hand were estimated via prosody of speech using openSMILE [64]. From these parameters gesture was matched from database. The three main categories of autonomous producing movements efforts are defined as rule-based methods, statistical models, and generative machine learning models. With the use of explicit phrase-to-gesture mappings, rule-based techniques severely restrict the diversity of gestures. Since sequences of gesture are constructed from pieces of actual motion data, statistical models can build realistic gesture forms by using the estimated conditional probability of certain speech variables happening in combination with a set of motion features. [65] Since the hands move so quickly and frequently experience motion blur during a typical sign language conversation, tracking the hands is a challenging undertaking. Hands are pliable things that can change posture and position. They block each other and the face, complicating techniques based on skin segments. Additionally, tracking may

be lost or the hands may become confused as they engage with one another. Early research required subjects to wear colored gloves, greatly simplifying the segmentation job. Acquisition mode, static versus dynamic signs, signing mode, single-handed versus double-handed signs, methods and techniques used, and the average accuracy are all classified as comparative parameters in this research and the systematic literature review strategy we used. These criteria have been used to analyze and document reviews of various sign languages, including American, Indian, Arabic, Chinese Sign Languages. and Ukrainian Sign Languages.[66]

The American Sign Language recognition of static words based on the Hough transform was provided by 300 examples of 20-sign images were gathered, and information such as the reference origin, shape orientation, and orthogonal scale factors were retrieved. The experimental findings demonstrated that the suggested technique is resistant in a significant way to changes in sign size, position, and orientation. To translate ASL words into English, A novel technique has been proposed in which created a recognition system that makes use of sensory gloves. They attained an accuracy of 90% by applying the neural network classification technique on static one-handed samples of 50 words. A novel sign language recognition system utilizing electromyography and arm sensors was suggested by Tangsuksant. They collected 40 American signs and than categorized them using Deep Learning algorithms including Naive Bayes, Nearest Neighbour, Decision Tree, and LibSVM, four classifiers. The experimental results and findings demonstrated that the SVM performs better than every other classification technique. A sign language recognition system utilizing Kinect and colored gloves was demonstrated by Usachokcharoen et al. [18]. From eight American signs, they retrieved depth, motion, and color aspects. It has been discovered that the system's accuracy is improved by the color feature [54].

2.7 Sign Language Processing (SLP)

Sign language is normally considered as gesture language. Uses of hand gestures are made instead of voice or audio for communication. This method is followed by those people who are abnormal or they may not have the ability to speak. For such people to communicate or interact with the modern technology such as computer have only one source as hand gestures or commonly called as sign language. [65] presents a techniques to generate gestures for an

artificial intelligent robot using Television shows especially their subtitles. Television shows were recorded with its subtitles which were extracted using OCR methods [66]. All the subtitles which were matched with the user entered word. This matching was performed by stemming approach [67] positive and negative sequences of subtitles were formed for any targeted word depending on the matching found. To extract the description of hands an articulated upper body gesture model was used. Various techniques are available for speech to gesture generation like in [35] the proposed system correlates gestures and prosody to build a probabilistic model. Stocks and holds of gestures were extracted from the motion capture data which are then clustered together for identification of motion sequence. Prosody based features were extracted from prosody of speech. These extracted features and stocks of gestures were combined to estimate a hidden Markov model. The body language was synthesized in an appropriate way for a live audio speech utterance. No appropriate validation technique could validate the method because no unique animation was found as correlation of body language and speech leads to many to may mapping a subjective evaluation was performed by conducting a survey.

F. Eyben et al. [68] introduced a novel framework for audio and emotion recognition. They have combined and made available multiple components in one package including speech recognition and file recording including audio. The proposed model was based on Real Time input. A novel study is introduced in [69] to analyze customers' perceptions on the basis of Natural Language Processing and Deep Learning Models. The NLP pipeline was designed in a systematic way like the input documentation was divided into multiple tokens then the tokens were arranged into sentences via proper sequence and afterward syntactic and semantic analysis was performed. Researchers have been working for years to improve and enhance text analysis in the context of NLP and introducing semantic analysis and other techniques [63] proposed a Novel approach and algorithm to improve semantic interpretation of the presented model have achieved better accuracy. Basically under the area of NLP scholars use to study the best way to use NLP in order to utilize users' reviews and their emotions and check what expressions the user is trying to give via them. [59]. This text analysis of user comments and expressions can be used to produce appropriate gestures. Before this modern Deep Learning Technology mostly work been found for producing gestures using hand-craft features with the use of typical sequence learning models and algorithms. For instance, For action recognition tasks human body data consisting of its skeletal features were given to the (HMM) Hidden Markov Model for processing and generating results.[60]

Camgoz et al [48] proposed a transformer based architecture for sign language recognition and translation via CLSR and SLT methods [69-71] and developed a baseline scenario to give direction for research. The objective was achieved by learning the probabilities $p(G/V)$ and $p(S/V)$ where G is the sign gloss sequence and S for spoken language sentences and V for a sign language video. The evaluation of this transformer architecture was done on challenging RWTP PHOENIX-Weather-2014T (PHOENIX14T) dataset [72]. Since last decennium the computer vision section has been analyzing and focusing on sign languages to attain the goal of building translation and production systems. [48, 73, 74] These systems have the ability to transform uttered sentences in a language like English into sign language and vice versa. Relation of these two entities is non monotonic because their order is not same and consequently they do not possess one-to-one mapping so they use CLSR methods for learning purposes.

In recent years Machine Learning and Deep Learning models have revolutionized every field, and various DL models are being presented for extracting visual features and performing sequence learning via encoding and decoding. Such models are promising to provide an effective environment for computer vision environments.[76] While performing sequence learning. The most popular Deep Learning model that has been used widely is (RNN) Recurrent Neural Network. RNN leads to some problems such as gradient descent vanishing issue and Long Term dependencies so to solve these problems (LSTM) Long Short Term Memory Networks was introduced which gave the idea of retraining Long Term Memory and save Long Term contextual information. Proposed a Temporal Convolutional Network to detect the location of action by calculating adjacent data points from heights and generating Short Temporary data for sequential input data. This model was used in translating Sign Language Videos.[75] For cases where multiple input sequence lengths vary from each other and to compensate for this problem CTC [54]was proposed which has been extensively used for voice and text recognition. Gesture generation and recognition tasks have widely been integrated into real-life applications to reduce human computer distance and provide an interactive and entertaining environment where advanced tools like chat gpt and Chabot have become the primary source of interaction with machines. The techniques behind generating such models include statistical modeling, processing of computer vision data and images, videos, and condensation algorithms that have also been integrated into Artificial gestures providing systems like Robots. [67] Large variety of raw data can be designed and modeled by using Deep

Learning and machine learning algorithms and model for creating movements specifically neural networks which can separately learn the correlation between voices that have high dimensions and targeted output of motion. They are able to produce unique motion that is a poor copy of the diversity of the training samples of data. The typical training methodology that used mean squared error loss produces averaged and smooth movements.[74]

A review of the possible encountered challenges while dealing with Sign Language is present in [75] regarding input modalities, dataset complications and models. Normally two input modalities are used for input to SLP model Visual and Lingual modality. For visual input deep learning algorithms such as Generative Adversarial Networks GANs, Long Short Term Memory LSTM networks are available. Recently Convolutional Neural Network CNN showed excellent performance in extracting features from an input image [76]. The other input modality text requires seq2seq models for processing such as Recurrent Neural Networks RNN. Processing a textual input is less complex than translation task. A specific domain is needed to be adopted for translation text because of unmatched words styles and a word poses different meaning in different languages. To cope with this challenge Transfer Learning Approach is used commonly by performing fine tuning in domain data for few epochs. Large amount of data is required for deep learning models for better generalization. This leads to a challenge of dealing with unseen data. Such words in data are tackled by bite pair encoding techniques such as Stemming. Few datasets are available publically such as RWTH-Phonex-2014T [77] and How2Sign [75]. RWTH-PHONIX-Weather 2014T [72] and How2Sign are used as benchmarks to evaluate the generalization of SLP models. Using these datasets and discussed input modalities with its applications a summary of Deep Learning SLP models is given in Table 2.1.

Table 2.1 Summary of Deep Learning SLP Models

Ref	Input Modality	Dataset	Description		
			Strength	**Weakness**	**Evaluation Parameter**
[33]	Text	Used a speaker specific dataset which was publically available in [32]	Produced Same quality gestures from text as produced from audio in	Low accuracy and produced gestures were specific to a speaker	PCK (0.288) MAE (0.958)
[32]	Audio	Own Dataset	Generated Good quality speaker specific gestures	Speaker-Specific gestures.	PCK (0.4) MAE (0.707)
[78]	RGB Video	ViSiCAST	Animation content was focused and compared with signers of human beings.	Non-manual features of human signers were not included.	Objective (58.4%) Subjective (58.6%)
[79]	RGB Video	Own Dataset	Naturalism was there in produced images and computational complexity was reduced.	Shoulder's position should be on the position of avatar's elbow instead of IK end effectors.	BLEU-4 (21.10) ROUGE (42.57)

[80]	RGB Video	Own Dataset	High viewer acceptance of the sign avatar improves understandability.	Bounded with in minor sign phases.	-
[48]	RGB video	PHOENIX-Weather 2014T	Strongly recognize sign videos.	Signs were not align in spatial domain.	BLEU-4 (21.80)
[81]	RGB Video	Own Dataset	Alignment of word order is strongly tuned with visual content in sentences.	Generalization on less datasets.	BLEU-1 (0.487) BLEU-2 (0.330)
[82]	Text	PHOENIX 14T	Training of the model was done via skeleton annotations which gave good results.	High Complexity.	
[83]	Text	PHOENIX 14T	Gloss information was not used.	High Complexity.	BLEU-4 (18.51) ROUGE (43.75)
[84]	Text	Czech news	Gives better results even skeleton parts are missing.	Facial expression may improve performance if added.	MSE (21.80)

2.8 Difference between Sign and Gesture

"I want that you all understand the difference between sign language and gestures. (.) Sign language is our own, it has been researched, it's related to our tradition, it's been linked to us for generations. That is sign language. Gestures are made by individuals, with imagination, to be able to communicate, to say something. (.) Gestures are based on impromptu thoughts, made up (.), created to make ourselves understood. Those are gestures. But language comes to us through generations." [85]

In an organized discussion by authors [85] on Difference between "Sign" and "Gesture" while explaining Sarita (a participant in the discussion) said these words clarifying the difference between Sign and gesture. She conformed universal outlook of Sign in Linguistics and Euro-American Deaf clique. Sign is a language which is commonly used by deaf people usually signs do not need any speech while gestures are co-speech and they are used by hearing people while they utter any sentence to improve the understandability of the listeners. The discussion [85] was not only verbally done but also practically. An example of gesture fig 2.6 and sign for "30" Fig 2.5 acted by Neeta (Assistant Teacher of deaf school children).

Figure 2.6 Sign of "30" using SIGN [85]

Signs are usually performed with the help of one hand as can be seen in Fig 2.5 representation of "30" is done with one hand in two steps.

Figure 2.7 Sign of "30" using GESTURE ($3 \times 10 = 30$) [85]

A Gesture is performed with both hands as in Fig 2.5 the Gesture of "30" will be completed in three steps because Gestures are more detailed so will require little more time than gestures to be performed. [86]

Furthermore, there exist languages based on Sign such as American Sign Language (ASL) and British Sign Language (BSL) therefore they are more specific than gestures and possess a particular meaning while Gestures are less specific as they require second entity Speech or Text for better understanding. More clearly in Fig 2.6 a gesture is presented which indicates something to drink but we can not specify whether it is a "water" or "bear" or "soft drink".

Figure 2.8 GESTURE of water [85]

In conclusion, Gestures are co-verbal and produced spontaneously while speaking and sign has its own specific and predefined meaning corresponding to any language such as ASL and BSL. It does not need for any other entity to convey its meaning yes.

2.9 Research Gaps and Directions

After an extensive survey of existing Literature of Gesture Generation Models multiple problems have been identified which led to development of text-to-gesture generation model discussed in coming chapters. Existing issues are related to model's complexity, Input Modalities with their scope and quality of generated output gestures. Some identified problems and existing gaps are listed below:

- Since Human Machine Interaction is the core of Artificial Intelligence the space between Virtual agents and Robots Specifically Humanoid robot with human is becoming narrow. This interaction is based on Speech till now. Considering text as a same worth entity on which this Interaction is based very little work is done on it and hence it is needed to be explored.
- Considering gestures as an easy way to communicate with Robots in the field of Artificial Intelligence multiple models have been proposed to generate gestures Artificial Gestures which are lacking in quality and needed to be more like human general.
- Most of the proposed models are based on Speaker-Specific data which limits its scope. Therefore, the Input Data is needed to be generalize like for Text Input Modality data should be general and same for audio.
- As Artificial Intelligence Models learn from data which is provided to them during phase, data should be sufficient but not less for training. Therefore, this problem needs attention of researchers.
- Recurrent Neural Networks (RNN) have showed good performance while dealing with sequential data. As producing motion with the help of individual gestures is all about sequencing the gestures. Therefore, it is required to be explored how a sequential model can affect the generated gestures sequence and quality.
- Section 2.8 clarifies the difference between Sign and Gesture. From the extensive Literature survey it has been observed that several approaches exit that produce sign from text input modality but text based gesture generation model are very rare.
- The available approaches of gesture generation model based on text input or audio undertake only standard English Language keywords. Though English language is understandable by most of the people still other languages are needed to be set as base for Artificial Intelligent Models.

- S2G Model is sensitive to noise as there are chances of noise in recorded audio for Model's training which may lead to unwanted Signals.

Several proposed Models have been proposed to generate gestures or sign. Such models are having deficiencies or have specific limitations. Like in S2G model proposed in [32] achieves an accuracy of 54 in term of Percentage of Corrected Key points (PCK) value and also the model is specific to speaker and in [33] an attempt has been made to produce same quality gestures from text using Convolutional Neural Network (CNN). Such targeting aspects and limitations convince researchers to explore more. This research is being conducted to target some of the above discussed challenges. This work is directed to propose a Text-to Gesture generation model which aims to improve the accuracy in terms of Percentage of Corrected Key-Points (PCK) value and gesture quality involving a sequential model. Also the problem of speaker specific based input modality either in textual form or audio speech is addressed in this research.

2.10 Summary

This chapter provides a solid foundation of research by providing an extensive Literature Survey on Gestures and its existing models. Critical Analysis of existing models related to multiple techniques is provided in this chapter. Moreover, Strengths and weakness of each technique is discussed theoretically. Finally, their deficiencies and limitations is discussed which should be undertaken as challenge for further research.

CHAPTER 3

METHODOLOGY

3.1 Overview

This chapter describes the research methodology used to design the text-to-Gesture model and also discusses the proposed hybrid deep learning model's architecture. The methodology followed in this research with all its phases is discussed in this chapter. What problems were undertaken to conduct this research are described in this chapter. This chapter expands the design phase and gives the overall view of the steps followed in this research. The aim of the proposed scheme is to generate quality gestures from text and improve the accuracy of the predicted gesture in terms of the Percentage of Corrected Key points (PCK) value by giving the recent publically available dataset. The use of an updated dataset is aimed at producing better and more significant features for the training of the model and choosing the correct methods of hyper-parameter tuning. This chapter discusses the requirements is elaborated. Data preprocessing reduced the complexity of data for extracting features and for embedding feature vectors. The proposed mechanism shows the complete architecture and strategy used to build the novel Hybrid Deep Learning model in order to meet the objectives to improve accuracy. Furthermore, the scope, size, and instances of the dataset are discussed in this chapter briefly.

3.2 Research Methodology

This section describes the methodology and operational framework of the proposed research work. The steps followed during research work and step-by-step procedures towards minimizing the distance with the objectives are discussed.

This Research consists of three phases as shown in Fig.3.1. The first phase is Analysis Phase. During the Analysis Phase, all the existing schemes and multiple solutions to produce gestures are studied and analyzed including the domain of Neural Networks, Morphemic Analysis, GANs, and several Data-base driven Approaches. Neural Networks are proven to give good results in complex Computations as they can deal with insufficient data and they can parallel process different problems. Morphemic Analysis has also produced better-quality gestures as it can recognize the meaning of a word up to its root meaning because gestures are synchronized with uttered speech or text. Generative Adversarial Networks (GANs) based approaches are available to generate gestures and they have shown the quality of output gestures as they have the ability to reproduce the data from instances. They are comprised of two neural networks which work in parallel and enhance parallel processing. Database-driven approaches are helpful in analyzing complex data and producing more than one result. Many gesture generation methods are proposed using data-driven techniques. From the available Literature, several gaps are identified and limitations are found and described in section 2.9 that still require focus ad research. By investigating these constraints, the problem statement is developed and objectives are set.

The second phase is the design and development phase which is considered the main one because this includes architecture and model building by analyzing all the requirements including Dataset and all the model parameters and Hyper-parameters. Analyzing that by giving text input better gestures can be produced the model is tuned to accept input in textual form and as the gestures and text sentences are synchronous and text is given in the form of word vectors in sequential order. Therefore, a sequential model is used to ameliorate the output gestures. Significant features are extracted from the input text using Convolutional Neural Network (CNN) and these drawn-out features are then fed into sequential Long Short Term Memory (LSTM) to generate 2-dimensional key points of gestures. To ensure sustainability and trustworthiness, the developed text-to-gesture model is tested upon multiple epochs and train test split.

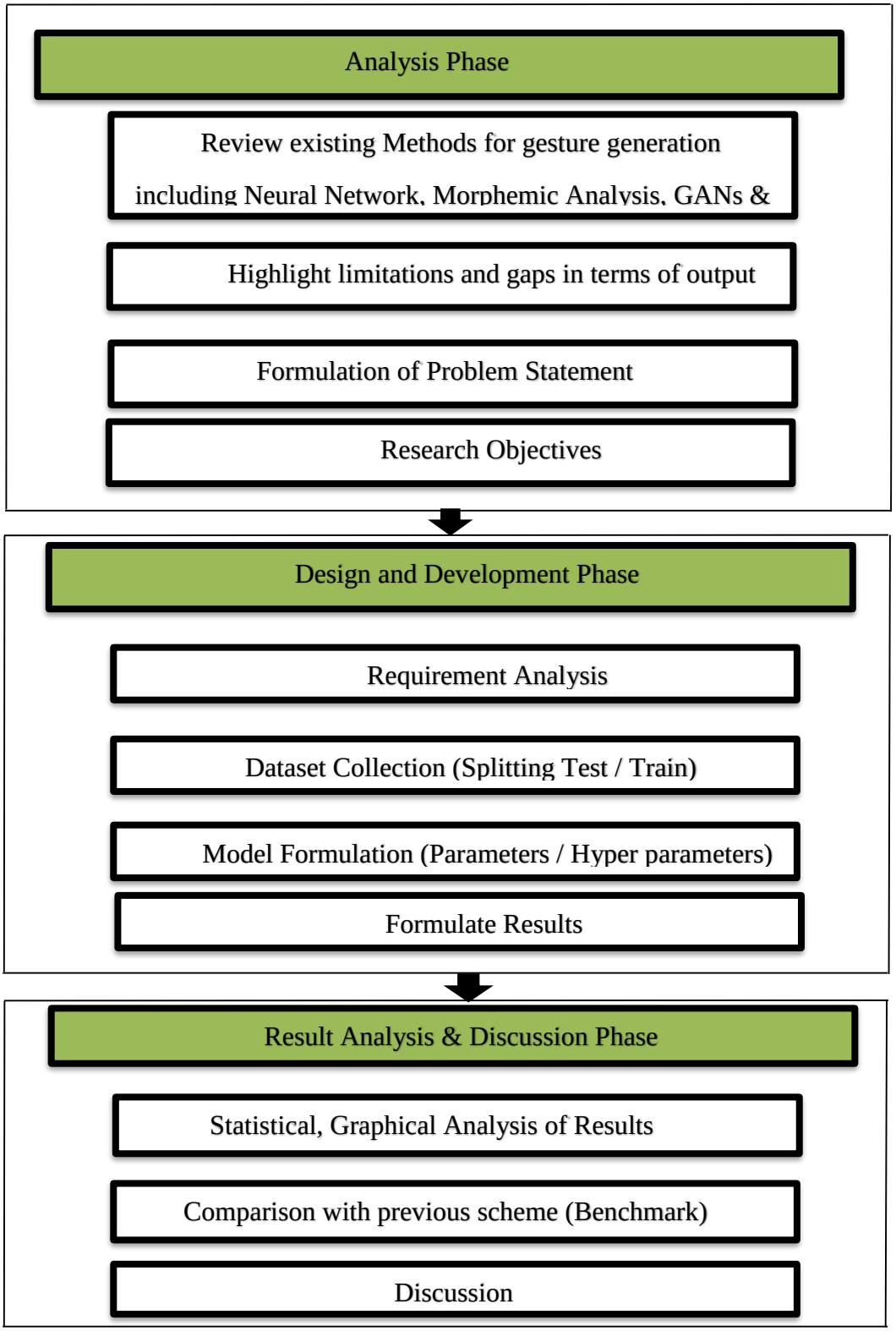

Figure 3.1 Framework of the Research Work

3.3 Requirement Analysis

Analyzing the requirements is the leading step to implement any proposed idea in which all the resources to be used can be analyzed and gathered. This step includes further sub-steps like gathering requirements, Prioritizing the requirements, Documenting requirements, validating requirements, and managing all the requirements. To implement the proposed text-to-gesture model, the first requirement is to have enough data, and a system with GPUs with several other dependencies is required to conduct the experiment which is discussed in detail in section 3.2.2.

3.3.1 Dataset

A speaker-specific text-to-gesture dataset is publically available in [33] and created in [32] to generate gestures and also used in this research. The original dataset consisted of 144-hour video data customized to study gestures and speech individually in a data-driven fashion of 10 individual speakers including 5 talk show hosts, 2 televangelists, and 3 lecturers. While recording the speakers faced the camera whose position was fixed samples can be seen in Fig. 3.2. 2 D key points were extracted from the videos using the “Open Pose” pose detection System. 49 key points were extracted corresponding to arms, hands, elbows, wrists, and shoulders.

Figure 3.2 Camera recording Speakers poses to extract key points. [32]

The actual size of the dataset was 411.61 GB as it further contained data of 10 speakers individually. A single speaker's data include the speaker's frames, Key-Points and words. Every speaker's word data set consisted of 23 different CSV files each containing an average of 4K instances and 2 columns where one column contains the specific extracted word and the other column contains the respective time and a "*BLANK"* keyword was assigned where there was no word against any particular frame as illustrated in Fig 3.3. 49 different Key Points were extracted in total and comprised of 2 dimension hence x coordinate and y coordinate of each key point was given.

839	<BLANK>	55.93333
840	<BLANK>	56
841	<BLANK>	56.06667
842	<BLANK>	56.13333
843	<BLANK>	56.2
844	<BLANK>	56.26667
845	<BLANK>	56.33333
846	<BLANK>	56.4
847	<BLANK>	56.46667

74	for	4.933333
75	you	5
76	you	5.066667
77	tonight	5.133333
78	show	5.2
79	show	5.266667
80	show	5.333333
81	show	5.4

Figure 3.3 Data Samples from csv files having extracted words against particular time frame.

3.3.2 System Requirements

System Requirements are the system capabilities that it should possess to run a specific software and to ensure that software is operated effectively. Usually these requirements include certain dependencies, Processor requirements, operating System, Storage space and Graphics Card etc. To implement the proposed system in this research required a number of requirements which are mentioned below:

- **Storage Space:** A space of 411.61 GBs was required to store the Key-Points, Frames of all 10 speakers. Shelly = 12.12 GBs, Seth = 85.04 GBs, Ellen = 29.71GBs, Conan = 50.31 GBs, Chemistry = 31.96 GBs, Angelica = 49.85GBs, Almaran 18.42 GBs.
- **Python 2.7:** Python 2.7 was introduced in 2010 which was the important release of Python. This python version is used to implement the proposed model in this research.

- **Cuda 9.0:** (Compute Unified Device Architecture) enables programmers to use the NVIDA GPU power for performing general purpose computing tasks including deep learning etc.
- **CuDNN:** Cuda Deep Neural Network Library allows to implement complex deep learning operations including pooling, normalization and implementation of various activation functions and perform functions accelerated to Graphical Processing Units. The advantage of cuDNN includes Pytorch, Caffe, NXNet frameworks.
- **Open CV:** (Open Source Computer Vision) This library is used to perform functions related to image processing such as to work with gestures frames for training purpose.

3.4 Pre-Processing

Pre-Processing refers to some operations, several techniques and approaches on raw data to prepare it for further use and according to the requirement. To make sure the data quality and effectiveness for data analysis to be performed some computations are needed to be made on data. In Deep Learning pre-processing is an important and essential step to be performed on Data including Normalization of data and to data according to the format for training of a Deep Learning Model and it is also promising in improving the accuracy of the model. This research is focused on general text as input so speaker information is therefore removed from the gestures which was provided to the model and taken as ground truth for training purpose. Dataset was comprised of 49 2 D key points corresponding to arms, hands, wrists, elbows and shoulders therefore, equivalent to 98 key points in total. Further, text for testing the performance semantic text was given as input which was taken from kaggle.com.

This data was used in [32,33] to train model on person's individual gesture styles whereas in the proposed Text-Gesture generation model the gestures are not supposed to be specific to any speaker so it was necessary to remove the speaker specific information from the dataset including gestures and words. In order to remove the speaker information which gives particulars of the speaker producing movements the procedure described in fig 3.4 was followed.

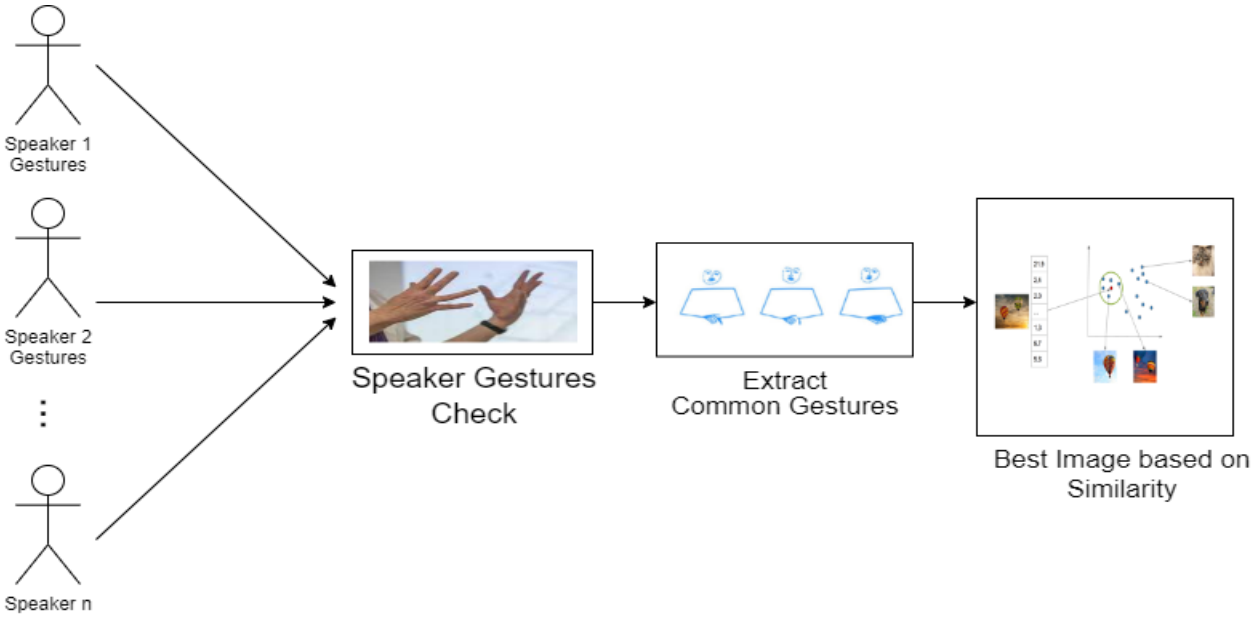

Figure 3.4 Pre-Processing

3.4.1 Checking Speaker's Gestures

Gestures Data was comprised of Gestures frames, words and Key-Points which were extracted from Open Pose consisting of 10 different speakers but the proposed research is upon generating gestures not particular to ant speaker. So to remove this speaker data from gestures and analyze the types of gestures and relevant Key-Points. The Key-Points were corresponding to shoulders, arms, wrists and hands.

3.4.2 Extracting Common Gestures

Gestures Dataset was comprised of words, Gestures Images and Key-Points. Words were extracted against a particular time frame and part has a specific gesture. As every speaker is either a T.V show host or a lecturer and have their own gesture images and words it was possibility that there would exist such words which are common in every speaker's data such as words used for Greetings, Introduction, Audience Interactions and to give recap of previous lecture or show etc. To remove the speaker information maximum number of common words were selected and against all those words the gesture of every speaker was selected. The experiment was performed on only three speaker's dataset Rock, Shelly and Almaram due to limited resources. Gathering common words for Rock there were 480041 words, for Shelly there were 1363701 and for Almaram there were 659058 words.

3.4.3 Selecting Best Gesture

In Words and Gestures dataset each word has at least 3 gestures as implementation of the proposed research is done on 3 speakers. In order to train the model on a particular gesture, Structural Similarity Index (SSIM) was used to find the best gesture image depending upon similarity value.

3.4.4 Structural Similarity Index

The Structural Similarity Index (SSIM) is a measure to check out resemblance in two images in image processing. One of the most profound measures in Deep Learning to find relation between images in terms of Similarity. It can also be implemented to evaluate Deep Learning Models which work for images. In addition, SSIM is also used as an evaluation measure in image restoration and generation. Structural Similarity Index (SSIM) captures the image similarity via three salient points which are Contrast, Structure and Luminance. In order to calculate the SSIM Index the deviation in these three key points is calculated and gathered via set of functions having weights. The resultant value is ranges between -1 and 1. If the value of resulting index is 0 it indicates there is nothing similar between two images if the value is 1 it indicates the input images are perfectly similar whereas -1 is the representation of complete dissimilarity. The similarity index between three gesture images of speakers was tested like first was compared with second than second was compared with third and third was compared with first and the gesture image with highest value i.e. more close to 1 was selected and respective key Points and words were undertaken for model training purpose.

3.4.5 Train Test Split

In order to approximate the performance of Machine Learning/ Deep Learning Algorithms the Train Test method is followed. Train Test Split is an easy procedure to be followed on data. The complete dataset is divided into two parts. One is training set which is utilized to train the Deep Learning Model and other is test set which is kept unseen from the model and through its results the performance of the model is evaluated. In this research the ratio 80:20 is set for train test split pf data where 80% of the data is used to train the model and 20% is set for testing purpose.

3.5 Proposed Text-to Gesture Model

In the proposed model 2 Deep Learning algorithms are put in to generate gestures. Deep Learning models are proven to give best results in multiple real life scenarios. In this research 2 most common deep learning algorithms Convolutional Neural Network and Long Short Term Memory network are used where Convolutional Neural Network is used to accept the text data as a sequence of input and converted into word vectors. 5 down block operations and 5 up block operations are performed to pass the time series information. [33] followed by a 3 layered Long Short Term Networks with a dense layer and flatten Layer. Finally, Gestures are generated using 98 Key-Points joint coordinates.

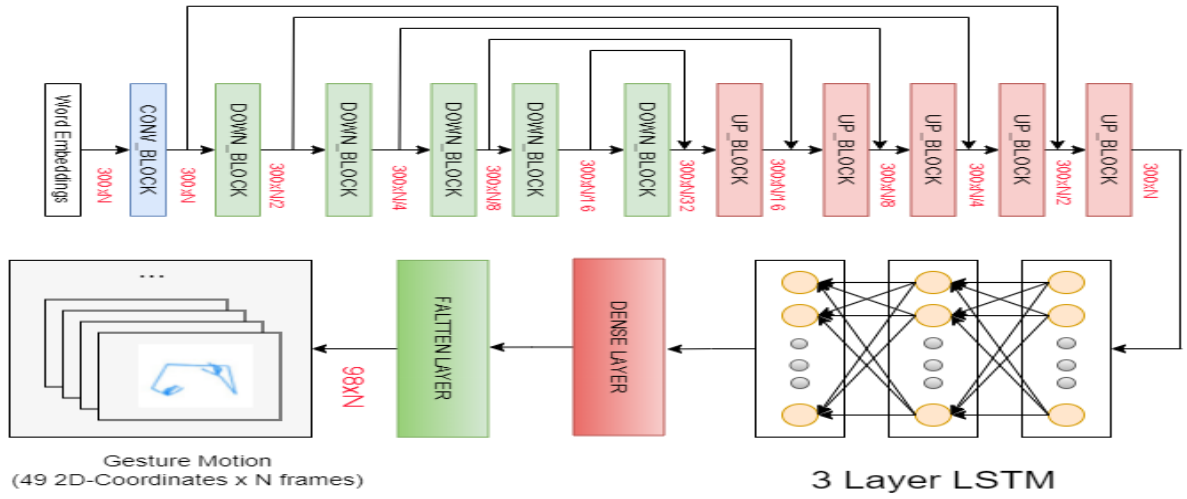

Figure 3.5 Proposed Model

3.5.1 Word Embedding Word2Vec

In Deep Learning and Machine Learning Models it is essential to convert the textual words into vectors while dealing with text and semantic data. When the words are represented into continuous vectors space the machine learning models can understand meaning and every context in an effective and constructive way. It also provides a mapping from high dimensions to lower dimensions. In the propose test to gesture model the technique Word2Vec is used to convert the textual words into vectors in which a single input word is transformed into real word vectors using one million English word vectors which were trained with Wikipedia subword 2017 with the help of FastText Python Library.

Word2Vec is a well-known model used for word embedding in Deep Learning which is specifically designed to learn a sequence of continuous vector representations based upon they occur in text. The presented Text to Gesture model is embedded using Word2Ve because its various impacts in the area of Text Classification, Sentimental Analysis and Machine Translation under Natural Language processing can be viewed [56].

3.5.2 Convolutional Neural Network (CNN)

Convolutional Neural Network is a vastly used algorithm in various Deep Learning and Machine Learning Models specifically when there comes any application or problem related to data visualization, classification of images and in images to detect any specific object or its edges. Inspired by the Virtual Cortex in animals which is comprised of quite a few layers of neurons, a convolutional Neural Network is also consisted of multiple layers like Pooling Layer, convolutional Layer. These layers are affixed together in a fully connected or partially connected way. CNNs have transfigured the grassland of computer vision and shown an exceptional performance in several tasks. In this proposed research, the advantageous uses of CNN have been taken out as following:

3.5.3 Down Sampling and Up Sampling

Down Sampling is an operation to minimize the structural directions of the feature map in Convolutional Neural Networks which is normally performed after convolutional layer. The purpose of this down sampling process is to minimize the numerical complexity of succeeding layers and keep track of the most significant features and the up sampling operations are performed to retrain the time series information. It is also known as Transposed Convolution or sometimes devolution and is the opposite of down sampling. The Down Sampling operations in Convolutional Neural Network are important in order to capture stratified representations and spatial resolutions in image data. In the proposed research the proposed architecture of convolutional neural network in is used which is comprised of 5 down block operations that are used to down sample the feature vector into 300 × N/32 where N represents the total number of frames in sequence which is given as input and further 5 up block operations are performed and each up block operation is used to add the time series information of the context of given input sequence which is a skip connection to transfer the time series particulars to decoder part.

3.5.4 3-Layred Long Short Term Memory Network (LSTM)

The 3-layered Long Short Term Memory Network (LSTM) architecture is the core of the proposed Text to Gesture Model. LSTM is well known for its high accuracy in processing Sequential Data like semantic text and video and audio processing. In this LSTM 3layered Network where each layer is put together at the top of each other. LSTM is a subtype of Recurrent Neural Network (RNN) and is enhanced version which is categorically structured to capture and safe long term dependencies. This deep Learing algorithm is widely used in recognition of Text and speech, Robotics and recognition of handwritten characters and digits. The architecture of LSTM Network is comprised of cells and every cell has its own three gates as input, forget and output gate. The architecture and parts of Long Short Term Memory Network used in the proposed Text to Gesture Generation Model is described below:

Input Gate: In order to modulate the information flow into the cell state of a Long Short Term Memory Network (LSTM) input gate is used which decides that significant part of input which should be kept in the current state and what part of it should be forgotten and which part should be retrained. This gate is comprised of Sigmoid Activation Function and receives data of previous states as input of current state. This input gate performs element wise multiplication in between candidate values which are obtained from current input and weights which are learned and sigmoid activation function's output. Addition operation is performed between the current state and produced result which allows LSTM Network to filter out information which is of no use or not have much importance and save relevant information. With the help of this process LSTM retrains long term dependencies.

Forget Gate: This gate plays a critical role in function of LSTM to forget the unnecessary information of the cell. It checks and decides what part of information is core of the input sequence and should be kept and other is discarded. The LSTM network adjusts its memory representation in a dynamic way and makes sure that to the point information in saved. This ability of LSTM is core in order to process sequential data.

Output Gate: Output gate of LSTM is responsible for the flow of necessary information from the current state to the next layer which determinates such parts of cell that should be used to produce the output. The output gate is comprised of Sigmoid Activation function which accepts the input data with all the information of preceding hidden layers and like previous layers this output layer also performs element wise product operation. This procedure allows LSTM network to choose information which is used to produce output.

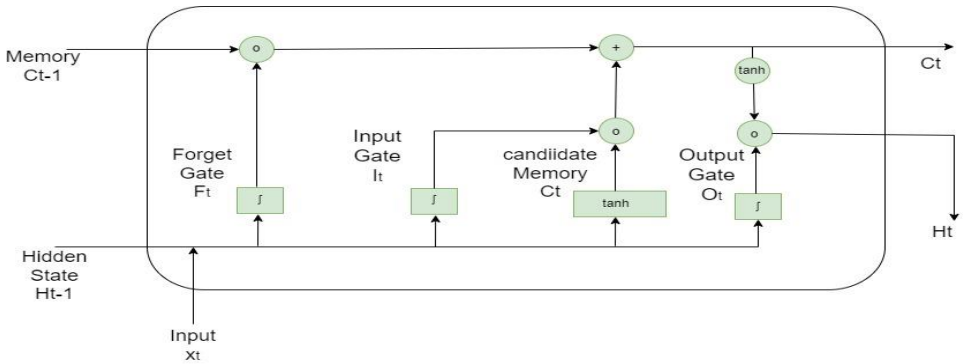

Figure 3.6 Architecture of LSTM

In the above figure 3.6 the architecture of LSTM used in Text to Gesture model is illustrated in which working of one cell with its Input, Forget and Output Gate is described. One cell of LSTM has 3 inputs x_t is input to the current cell, H_{t-1} represents the input coming from the previous hidden state and C_{t-1} which is a Long Term Memory Input whereas C_t and H_t are the outputs going out of the current state. The following equations represent the LSTM.

$$I_t = \sigma\,(w_i[h_{t-1}, x_t] + b_i) \qquad \text{(i)}$$

$$F_t = \sigma\,(w_f[h_{t-1}, x_t] + b_f) \qquad \text{(ii)}$$

$$O_t = \sigma\,(w_o[h_{t-1}, x_t] + b_o) \qquad \text{(iii)}$$

In the above mentioned equations I_t, F_t and O_t are the Input, Forget and Output gates respectively. σ is the Sigmoid Activation function used in LSTM, w_x represent the weights of neurons, h_{t-1} is the output coming from the previous cell of Ling Short Term Memory Network, x_t is the input at time t and b_x is the bias input.

3.5.5 Sigmoid and Tanh Activation Functions in LSTM

In order to introduce the non-linearity and to control the information flow activation functions are used in LSTM. This network uses two activation functions Sigmoid and Tanh. Sigmoid is used in forget, input and output gate. A real get and maps these values between 0 and 1 defined as $\sigma(x) = \frac{1}{1+e^{-x}}$

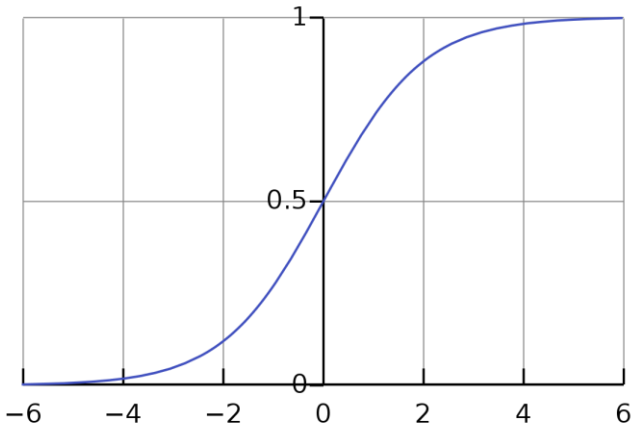

Figure 3.7 Sigmoid Activation Function

The hyperbolic tangent function activation is responsible for introducing non-linearity in hidden and cell state of an LSTM network. It maps the input value of real value between -1 and 1. Tanh has the capability to process both positive and negative values. It is defined as $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

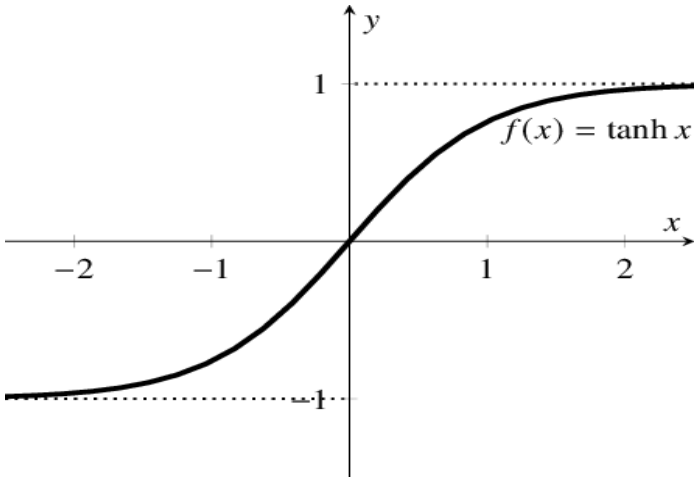

Figure 3.8 Tanh Activation Function

While processing sequential data both tanh and sigmoid activation functions play their role to retrain the selective information, learn complex patterns and to introduce the non-linearity in LSTM output and maintain the Long and Short Term Memory where Short Term Memory is maintained within a single cell and Long Term Memory is maintained over the whole LSTM network.

3.5.6 Dense Layer

With a 3-Layered LSTM network model a Dense Layer is used in Text-to-Gesture Model in this proposed research. Each neuron in this dense layer is affiliated with each and every other neuron of previous layer. In this layer a multiplication operation is performed in between the weights and input coming from the previous layer values produces after this operation into activation function in order to produce non-linearity in the final output layer also in this layer complex patterns of sequential patterns are learned via performing adjustment of bias and weight values. The neurons used in this layer represent the number in the output layer.

3.5.7 Flatten Layer

A flatten layer is commonly a transition layer which is used to convert multi-dimensions of data into a single dimensional or a flat vector in a neural network model. It is used to change the shape of output of neural network using parameters such as width, depth and height etc and transforms than into single 1-D vector. Flatten layer does not change the values of output or performs any complex operation its task is simply to resize and change the shape of output of neural network. In the proposed text to gesture model this flatten layer is used to map the output into 98 2-D of gestures to produce a sequence of gesture according to the number of input sequence words N.

3.5.8 Generating Gestures

In the proposed Text to Gesture model 98 2-D key-points joint coordinates are used to train the model relevant to arms, hands, wrists and shoulders. The model performed the following mapping $\boldsymbol{G : R^{300\times N} \rightarrow R^{98\times N}}$ where 300 is the number of dimensions of input vector, 98 is the specific number of gesture used and N represents the number of words in input sequence. Model accepts a sequence of text as input and generates a gesture image against word via correctly predicting key Points.

3.6 Adam Optimizer

Deep Learning and Machine learning usually process huge amount of high dimensional data. In that case fine tuning of the parameters and hyper-parameters of the model is essential. To perform these task optimizers are used. AN optimization algorithm targets two things loss and accuracy via fitting the best values as model parameters and hyper-parameters. In the proposed Text to Gesture model Adam (Adaptive Moment Estimation) is used which combines features of AdaGrad and RMSProp algorithm is used as optimization algorithm that specifically deals with large data set and choose optimal learning rates. Adam optimizer first initializes the weights than calculates gradients with the help of back propagation algorithm with respect to loss function. Than moving average of gradient and bias correlated average is computed and finally weights are updated and repeats this step until desired output is reached.

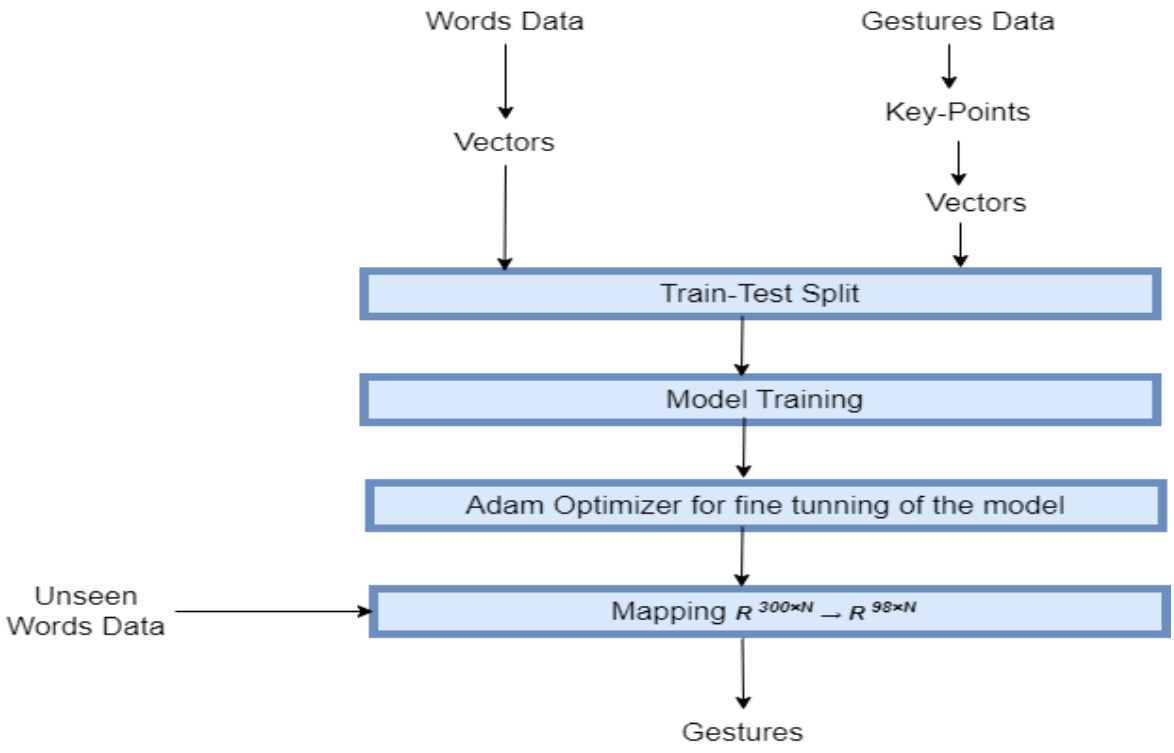

Figure 3.9 Text to Gesture Generation flow

3.7 Summary

This chapter describes the details of techniques and steps taken to generate Text to Gesture model. At first, the strategy how to remove speaker information from the gestures in

the available data set is explained. Moreover, what algorithms are used in the formation of model in order to deal with sequential data as the data being dealing with was large in number. Apart from this activation functions and optimization function algorithms used are elaborated in brief. Finally, the flow of Text to Gesture Generation is elaborated and explained diagrammatically.

CHAPTER 4

RESULTS AND ANALYSIS

4.1 Overview

In this chapter, the implementation results of the proposed Text to gesture model and techniques presented in Chapter 3 will be discussed. Integrating the sequential algorithm in Gesture Model is core of the proposed research. What influence it gave on gesture quality will be discussed in detail in this chapter furthermore this chapter describes the implementation platform, experimental settings, and parameters used to evaluate the performance of the model. The organization of this chapter is like in Section 4.2 gives details about evaluation parameters and obtained results, and analysis and discussion are described in Section 4.3. The significant achievement and comparison with other proposed gesture models is illustrated in Section 4.4. Finally, this chapter is summarized in Section 4.5.

4.2 Evaluation Parameters

An evaluation parameter has paramount value in any Deep Learning and Machine Learning Model because it is in service of a qualitative measure and estimates the performance of a trained deep learning model. An evaluation parameter not only provides a way to judge the model's performance but it also gives a comparison study with other state-of-the-art proposed models which is an effective way to evaluate a model. The proposed Text to Gesture model is evaluated on the basis of different parameters described in the sections below:

4.2.1 Percentage of Corrected Key Points (PCK):

PCK is used as a parameter to judge the performance of Artificial Intelligent models specifically designed for applications related to Computer Vision, estimating the pose of humans or any moving object. In such applications model usually works on Key-Points extracted from the image or any moving object in Real Time Scenarios.[33] It gives the accuracy of the model's prediction with respect to Key-Points of Ground Truth data. PCK performed an evaluation of the proposed Text to Gesture following some steps.

- **Defining a Threshold Interval:** It defines a threshold distance in order to estimate a pose that dictates a specific value of the Key-Point that can be considered as correct which is required to be close to the Ground Truth.
- **Computing accurate Key-Points:** For each value of Key-Point that is predicted by the model the distance between this predicted Key-Point and the actual ground Truth is calculated. If the resultant value lies in between the defined interval, then the predicted Key-Point is considered as accurate.
- **Calculating Percentage of the Correctly Predicted Key-Points:** In order to evaluate the model's accuracy, the next step is to calculate the percentage value of all the corrected Key-Points is calculated out of total extracted Key-Points. This percentage value is the actual calculated PCK representing the efficiency of the model corresponding to Ground Truth.

While calculating a PCK value defining a correct threshold is a very important task and is a very sensitive step toward accuracy as it defines the specific area upto which key Points are said to be correctly predicted. Calculating PCK of the proposed model the set threshold was moderate and also different thresholds were tested for the proposed model to authenticate the performance. Its biggest advantage was that only those Key-Points were considered correct which came in between a narrow threshold and so close to the ground truth. Setting up a narrow and close threshold gives higher precision [70] The threshold values used for the evaluation of the proposed model were $\alpha = 0.1, 0.2$. [33]

Correct Key-Points = No. of Key Points In between Threshold

$$PCK = \frac{No\ .of\ Corrected\ KeyPoints}{Total\ no.of\ KeyPoints} * 100$$

4.2.2 Mean Absolute Error (MAE)

The second parameter used to authenticate the proposed model's performance is Mean Absolute Error (MAE) which is the most commonly used metric in order to judge how well Deep Learning and Machine Learning model performs. MAE shows the best results in applications where there are regression tasks. It calculates absolute gaps between ground truth or actual data set and predicted values. The advantage of MAE is that it is not sensitive to outliers and it also gives a view of the model's progress insights as it gives an average of individually calculated absolute errors. The mathematical formula to calculate MAE is given as:

$$MAE = \left(\frac{1}{n}\right) * \sum |y_{actual} - y_{predicted}|$$

In the above equation, MAE denotes the value of the calculated error, n is the total number of instances in the available entire dataset, here actual and predicted are the original and predicted values of any marked variable or instance respectively. While computing MAE for the proposed Gesture Model following steps were performed.

- For each instance in the dataset the actual values were taken out.
- Against every data point absolute gap between the actual and the model's predicted value was calculated.
- All the resultant values were added and computed as the average.

To analyze the MAE value to evaluate the model it should be considered that the minimum MAE value is considered better which indicates that the produced or predicted value by the model is closer to actual ground truth values and it also indicates the unit of number is the same as the target variable in real-time scenarios.

4.3 Experimental Settings

During the experimentation and implementation phase following are some important settings used for this experimenting research.

Table 4.1 Experimental Settings

Setting	Value	Description
Batch Size	32	A batch size of 32 is used in experimentation, indicating that the proposed model's parameters were updated after every 32 samples.
Device	Cuda	The proposed model was trained on System having Graphical Processing Unit (GPU) having the Cuda framework. Cuda provides a parallel Computing Platform.
Epochs	90	100 Epochs were used for fine training of the model which means the entire dataset was passed 50 times through the model for training. Hence it took much time to train the model on such large data described in Chapter 3.
Model save Interval	10	Setting the interval of the model as 10 says that the weights and bias values during training are saved on disk after each 10^{th} iteration.

4.4 Results and Discussion

The proposed Text Gesture Model was evaluated using two performance evaluation matrices (PCK) Percentage of corrected Key-Points and Mean Absolute Error (MAE). Model implementation was performed in the PyTorch environment. Figure 4.2 illustrates the maximum Percentage we achieved for correctly predicted key points with respect to multiple epochs respectively. For PCK two thresholds α = 0.2, 0.3 have been used to evaluate performance from multiple views. 0.2 can be considered as a tight threshold and 0.3 a little moderate. Our dataset consisted of key points related to hands, arms, wrists, and shoulders. All joints are close to each other so setting a tight threshold may not create any ambiguity for achieving high performance. The Graph shows an increase in the value of PCK against the number of training epochs which illustrates the efficiency of the model.

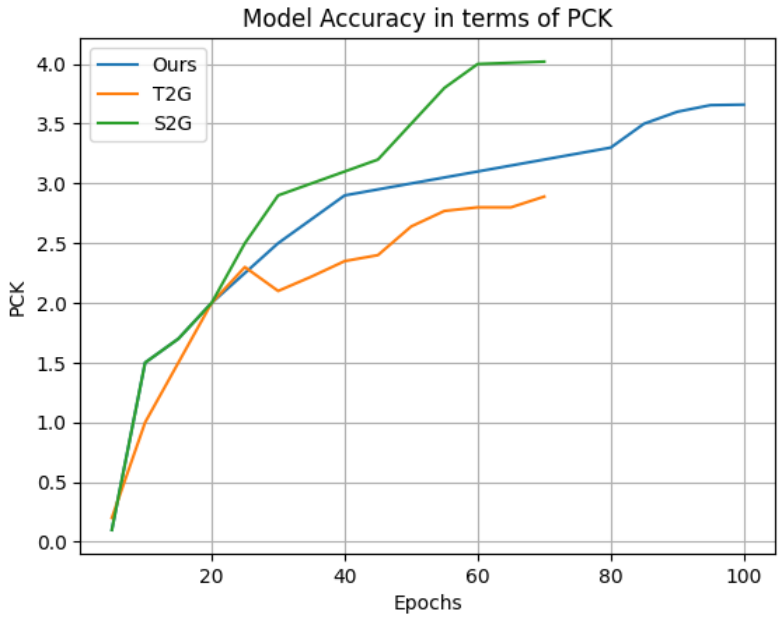

Figure 4.1 PCK

The highest value achieved and correctly predicted joint coordinates predicted by the model in terms of percentage is 0.35, more than the base model in [33] demonstrates the significance of the sequential 3-layered LSTM model.

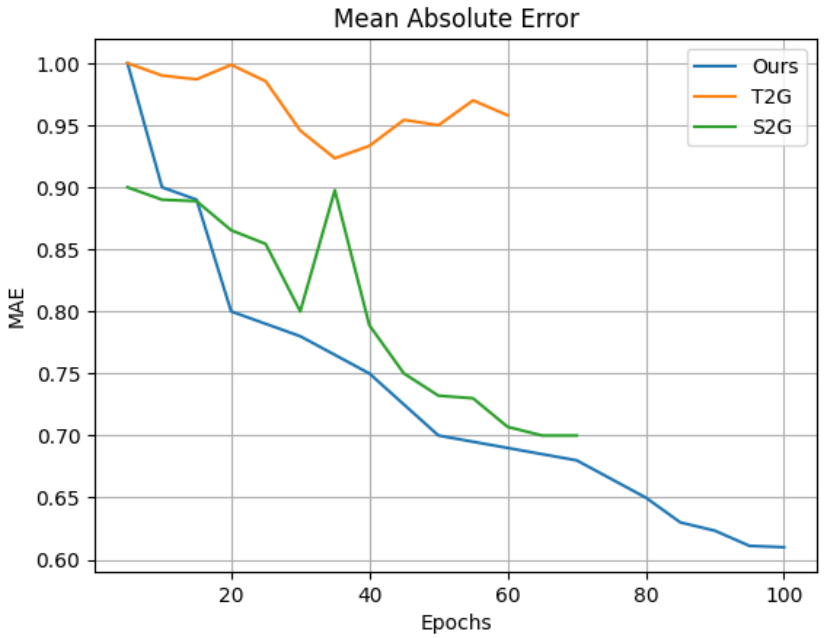

Figure 4.2 MAE

The above figure shows the fall in Mean Absolute Error (MAE) which says that the proposed Gesture model has decreased the rate of error in predictions.

4.5 Benchmark Dataset

While evaluating model's performance datasets play a crucial role in doing so. A benchmark dataset is widely used and accepted data that have been used already for available DL/ML models training and evaluating performance. While comparing two different models a common dataset owns significant value in providing just and fair comparison and results. It also highlights the positive and negative points of each model tested on a specific dataset as it provides annotated ground truth. The proposed Gesture model is evaluated on the publically available gesture dataset in [32] also used in [33]. The description of this dataset is presented in previous chapter. The following table clearly describes the comparison of the proposed text-to-gesture model with previous models.

Table 4.2 Comparison of Proposed Hybrid DL Text to Gesture Model

	PCK	**MAE**	**Threshold**	**Speaker Specific**	**Standard Deviation**
Our Hybrid Model	0.36	0.6	α = 0.2, 0.3	No	0.723
T2G	0.288	0.958	α = 0.1, 0.2	Yes	0.896
S2G	0.4	0.707	α = 0.1, 0.2	Yes	0.622

the above table, it can be seen that the proposed model has shown better results as compared to other models and improved the gesture quality in terms of PCK, and reduced the error in predicting key points. All the models used the same benchmark dataset [32,33] which contained the gestures of ten speakers specifically T.V hosts and lecturers along with their textual words respectively. All the persons were said to face the camera while delivering lectures. Key-Points were extracted using OpenCV [75] words were extracted from videos using Cloud Speech to Text in [33]. S2G [32] is based on voice input and generates gestures from speech. T2G [33] used the same gesture data and extracted words from video speech. The same data have been used in this research and removed speaker information.

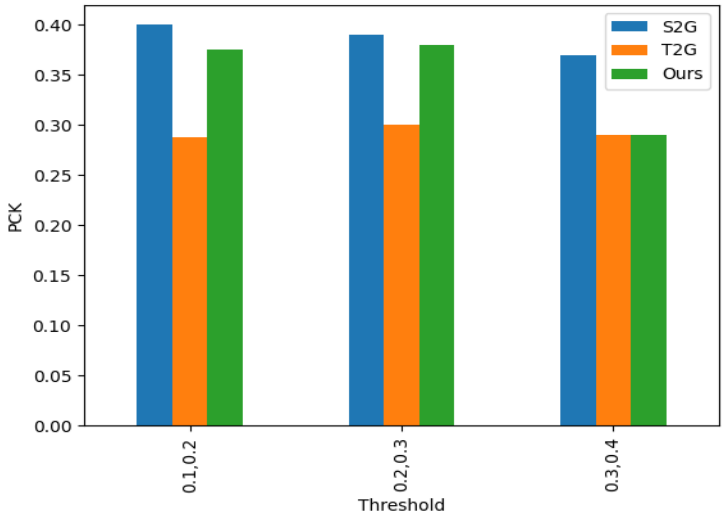

Figure 4.3 Comparing PCK upon threshold

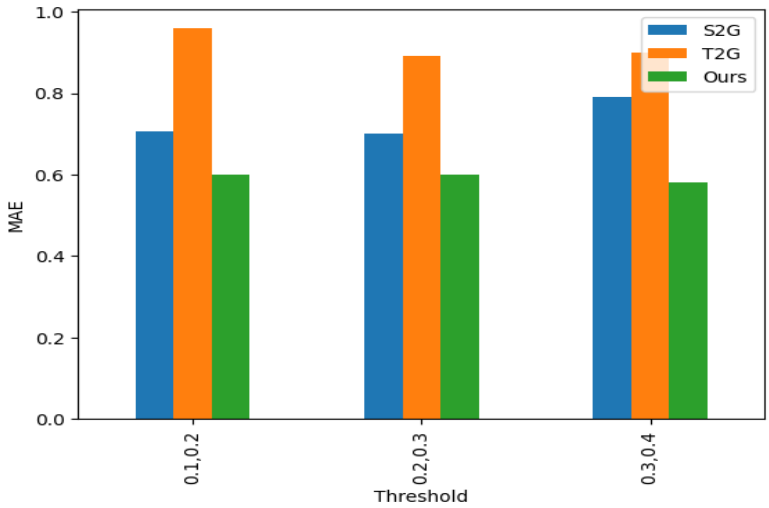

Figure 4.4 Comparing MAE upon threshold

4.6 Summary

In this chapter, the experiments and implementation of the proposed research in this thesis have been discussed. The experimental phase was performed in the PyTorch environment using the device Cuda. The proposed Hybrid Model was tested on a benchmark dataset and showed better performance as compared to other Gesture Models.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Overview

The principal emphasis of this research was mainly to investigate the problem of gesture accuracy and speaker specific movements and provide with a solution in the form of a deep learning model and techniques that can generate gestures independent of any speaker and improved accuracy in terms of PCK value. According to our knowledge the proposed model is first in its nature that generates hand movements independent of any person and against a text input modality. The proposed Text to Gesture model is to magnify the previous gestures and enhance them with new techniques and methodology. In this chapter complete summary of the research is provided more briefly in Section 5.2 the significant contributions of this research are summarized Section 5.3 describes certain issues and the directions that can be taken to further enhance the research. The center of this thesis is to design and develop a Text to Gesture generation model that can provide improved accuracy via PCK value. Furthermore, the influence of the proposed model is inspected by performing comparison with the available state of the art techniques.

5.2 Summary of research contribution

The proposed Text to Gesture model has enhanced the gesture model [33] in following significant ways:

i. **Gestures are independent of any speaker:** The proposed model presented in this research thesis is free of any person or speaker.

ii. The available techniques used data that was specific to some speakers in order to perform model training and produce gestures. This improves the chances of the model's limitation. To avoid this issue and maximize the model's flexibility the techniques have been used in the proposed model in order to remove the speaker information from the dataset. Figure 3.4 illustrates the strategy used to remove speaker information from the data used in [33] contained gestures of the ten persons. Gestures against the most common words are extracted and on the basis of image similarity best image against a word is selected and a dataset is formed containing words gestures data.

iii. **High Accuracy:** Accuracy is an important outcome of Deep Learning and machine models which tells the efficiency and performance of the model. The proposed gesture generation model is based on the Hybrid Deep learning approach and comprised of two major Deep Learning Algorithms Convolutional Neural Network and Long Short Term Memory Network its comparison with other state-of-the-art techniques producing movements clearly highlighted the accuracy achieved by the presented model in this research thesis. The proposed Model has increased PCK value by up to 10% and has minimized the Mean Absolute Error (MAE). The influence of using a sequential Deep Learning Algorithm LSTM on results can clearly be seen in sections of the previous chapter.

iv. **Gestures against Text Input:** Text owns supreme importance in various fields including documentation, sharing of information and knowledge. Hand movements play a critical role in providing interaction between natural and artificial entities including humans and Computers. Producing hand movements from text is therefore essential for multiple real-life applications. Gestures against words provide an immersive way to provide interaction between the user and the computer. Available models and techniques producing gestures are based on voice input [32] such schemes are mostly not benefited by many advantages. Therefore, the presented model in this research is based on text input and produced quality gestures.

Artificial Intelligence and Deep Learning techniques are promising in minimizing the difference between human and Machines therefore, the proposed text-to-gesture model is based on Hybrid Deep Learning Approach and have shown the best results as compared to other techniques [32,33] There is no doubt much work has been done in this research but still there exist various multiple ways in which this research can explored.

5.3 Applications

Artificial Intelligence has a lot of Applications in any field of real life. The proposed research owns the domain of Artificial Intelligence so it can be deployed in multiple real-life scenarios. Some of them are listed below:

i. **Translating Sign Language:** A Text to Gesture can be used for providing assistance to hard-of-hearing groups and an inclusive environment can be made available for deaf people. This model provides for implementation for translating text into signs.
ii. **Visual characters and Assistants:** Gesture models have a vast variety of applications in Virtual reality (VR) and Augmented Reality (VR) Environments for training avatars and animated characters to correspond against the text. [84,85] This can provide maximum interaction of humans with virtual assistants.
iii. **Robotics:** In order to provide an effective Human-Computer interaction the gesture model can be deployed and integrated into Artificial Robots for the interpretation of textual commands and use appropriate gestures against every word.
iv. **Modern Assistive Technological Systems:** People survive assistive devices which acts like personal assistant for them, such devices can be integrated with a gesture generation model various text messages can be translated into gestures to contact with the environment in a natural way.
v. **Healthcare:** Gesture models can be integrated and embedded in rehabilitation exercises and provide immediate guidance by performing movements accurately.
vi. **Entertainment:** Text to Gesture models can be integrated in storytelling applications such as games and other applications providing enjoyment and help little learners to grow up in an effective and interactive environment.
vii. **Artificial Intelligence-based Interfaces:** Using sentiment analysis on the context and conversations between Artificial Intelligent tool like chat-gpt and the user and provide expressive use of AI technology.
viii. **Story Telling:** Artificial Gestures can be used in story telling where story books can be used to give input in the form of text.

5.4 Limitation

The presented Gesture Generation model in this research thesis is although producing quality gestures yet it is under certain limitations. The dataset used for training the model consisted of some limited gesture features including key points of wrists, arms, hands, and shoulders also the input text is containing words for only English Language. These limitations can lead to explore more directions for further research.

5.5 Future Work

There is no limitation in the area of research. Artificial intelligence is the only field of Computer Science which is providing multiple ways and areas for researchers and scholars to explore. The presented research in this thesis can also be expanded in multiple directions to improve the gesture quality and performance of the model. Some of them are discussed below:

i. **Language:** The proposed model is able to produce gestures against text input modality which is based on English Language. This work can be enhanced by giving the model input words of any other language other than English. This can be achieved by training the model on that specific language characters' dataset.[80,81] It should also be noticed that gestures of different languages may vary from each other.

ii. **Including Facial Expressions with Hand Gestures:** The Key-Points of gestures on which the Proposed Gesture Model is trained in based on some limited features as it includes Key-Points relevant to arms, wrists, hands, and shoulders. This provides a direction to improve the gesture model and include facial expressions as well in the dataset. For Example: when there comes any positive word and manipulating the context of the word and performing some Natural Language Processing task the movement for that specific text according to facial smile can be generated. [88] Some researchers have proposed techniques in this concern.

iii. **Performance Evaluation Measures:** The proposed Text to Gesture Model is evaluated on the Percentage of Corrected Key-Points (PCK) to view its performance as PCK is the most commonly used performance evaluation measure for joints Key-Points data. In order to enhance research in this direction more performance measures can be generated and developed to evaluate the performance of a gesture model. This will increase multi-model performance evaluation.

iv. **Real-Time Gestures:** In the introduced Gesture model in this research the training of the model is performed and gestures are generated on a secondary dataset. This gives a unique direction for research and models can be designed and developed to produce gestures against real-time data. [87] This conduct is already under analysis therefore several initiatives have been taken upon this direction.

.

REFERENCES

1. Cassell, J., *Embodied conversational agents: a new paradigm for the study of gesture and for human-computer interface.* Gesture, speech, and sign, 1999: p. 203-222.
2. Kendon, A., *Gesture: Visible action as utterance*. 2004: Cambridge University Press.
3. Cassell, J., D. McNeill, and K.-E. McCullough, *Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information.* Pragmatics & cognition, 1999. **7**(1): p. 1-34.
4. André, E. and C. Pelachaud, *Interacting with embodied conversational agents.* Speech technology: Theory and applications, 2010: p. 123-149.
5. Bickmore, T.W., et al., *Usability of conversational agents by patients with inadequate health literacy: evidence from two clinical trials.* Journal of health communication, 2010. **15**(S2): p. 197-210.
6. Mayer, R.E. and C.S. DaPra, *An embodiment effect in computer-based learning with animated pedagogical agents.* Journal of Experimental Psychology: Applied, 2012. **18**(3): p. 239.
7. Salem, M., et al., *To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability.* International Journal of Social Robotics, 2013. **5**: p. 313-323.
8. Iverson, J.M. and S. Goldin-Meadow, *Why people gesture when they speak.* Nature, 1998. **396**(6708): p. 228-228.
9. Tang, Z., et al., *MLP-JCG: Multi-Layer Perceptron with Joint-Coordinate Gating for Efficient 3D Human Pose Estimation.* IEEE Transactions on Multimedia, 2023.
10. Wang, J., et al. *A portable artificial robotic hand controlled by EMG signal using ANN classifier*. in *2015 IEEE International Conference on Information and Automation*. 2015. IEEE.
11. Fang, H.-S., et al. *Learning pose grammar to encode human body configuration for 3d pose estimation*. in *Proceedings of the AAAI conference on artificial intelligence*. 2018.
12. Ionescu, C., et al., *Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments.* IEEE transactions on pattern analysis and machine intelligence, 2013. **36**(7): p. 1325-1339.
13. Poppe, R. *Evaluating example-based pose estimation: Experiments on the humaneva sets*. in *CVPR 2nd Workshop on Evaluation of Articulated Human Motion and Pose Estimation*. 2007. Citeseer.
14. Sharma, V., et al., *Optimizations of parameters for quality spawn production.* Mushroom Research, 2013. **22**(1): p. 31-36.
15. Nishikawa, A., et al., *FAce MOUSe: A novel human-machine interface for controlling the position of a laparoscope.* IEEE Transactions on Robotics and Automation, 2003. **19**(5): p. 825-841.
16. Becker, D.A. and A. Pentland. *Staying alive: A virtual reality visualization tool for cancer patients*. in *Proceedings of the AAAI*. 1996.
17. Wachs, J.P., et al., *A gesture-based tool for sterile browsing of radiology images.* Journal of the American Medical Informatics Association, 2008. **15**(3): p. 321-323.
18. Lukowicz, P., et al., *Wearit@ work: Toward real-world industrial wearable computing.* IEEE Pervasive Computing, 2007. **6**(4): p. 8-13.

19. Starner, T., et al. *The gesture pendant: A self-illuminating, wearable, infrared computer vision system for home automation control and medical monitoring*. in *Digest of Papers. Fourth International Symposium on Wearable Computers*. 2000. IEEE.
20. Gutiérrez, M., et al. *Telerehabilitation: Controlling haptic virtual environments through handheld interfaces*. in *Proceedings of the ACM symposium on Virtual reality software and technology*. 2004.
21. Patel, R. and D. Roy. *Teachable interfaces for individuals with dysarthric speech and severe physical disabilities*. in *Proceedings of the AAAI Workshop on Integrating Artificial Intelligence and Assistive Technology*. 1998.
22. Starner, T., et al. *MIND-WARPING: Towards creating a compelling collaborative augmented reality game*. in *Proceedings of the 5th international conference on Intelligent user interfaces*. 2000.
23. Freeman, W.T. and M. Roth. *Orientation histograms for hand gesture recognition*. in *International workshop on automatic face and gesture recognition*. 1995. Citeseer.
24. Wachs, J.P., et al., *Vision-based hand-gesture applications.* Communications of the ACM, 2011. **54**(2): p. 60-71.
25. Cook, K.A. and J.J. Thomas, *Illuminating the path: The research and development agenda for visual analytics*. 2005, Pacific Northwest National Lab.(PNNL), Richland, WA (United States).
26. Rogalla, O., et al. *Using gesture and speech control for commanding a robot assistant*. in *Proceedings. 11th IEEE International Workshop on Robot and Human Interactive Communication*. 2002. IEEE.
27. Hasanuzzaman, M., et al. *Real-time vision-based gesture recognition for human robot interaction*. in *2004 IEEE International Conference on Robotics and Biomimetics*. 2004. IEEE.
28. Yin, X. and X. Zhu. *Hand posture recognition in gesture-based human-robot interaction*. in *2006 1ST IEEE Conference on Industrial Electronics and Applications*. 2006. IEEE.
29. Sheridan, T.B. and W.R. Ferrell, *Remote manipulative control with transmission delay.* IEEE Transactions on Human Factors in Electronics, 1963(1): p. 25-29.
30. Nielsen, M., et al. *A procedure for developing intuitive and ergonomic gesture interfaces for HCI*. in *Gesture-Based Communication in Human-Computer Interaction: 5th International Gesture Workshop, GW 2003, Genova, Italy, April 15-17, 2003, Selected Revised Papers 5*. 2004. Springer.
31. Minnen, D., et al., *Performance metrics and evaluation issues for continuous activity recognition.* Performance Metrics for Intelligent Systems, 2006. **4**: p. 141-148.
32. Ginosar, S., et al. *Learning individual styles of conversational gesture*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
33. Asakawa, E., et al., *Evaluation of text-to-gesture generation model using convolutional neural network.* Neural Networks, 2022. **151**: p. 365-375.
34. Taylor, P. and A. Isard, *SSML: A speech synthesis markup language.* Speech communication, 1997. **21**(1-2): p. 123-133.
35. Levine, S., C. Theobalt, and V. Koltun, *Real-time prosody-driven synthesis of body language*, in *ACM SIGGRAPH Asia 2009 papers*. 2009. p. 1-10.
36. Ahuja, C., et al. *Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach*. in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. 2020. Springer.

37. Wu, B., et al., *Modeling the conditional distribution of co-speech upper body gesture jointly using conditional-GAN and unrolled-GAN.* Electronics, 2021. **10**(3): p. 228.
38. Chiu, C.-C. and S. Marsella. *How to train your avatar: A data driven approach to gesture generation*. in *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11*. 2011. Springer.
39. Alemi, O., W. Li, and P. Pasquier. *Affect-expressive movement generation with factored conditional restricted boltzmann machines*. in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2015. IEEE.
40. McNeill, D., *Hand and Mind: What Gestures Reveal about Thought Univ*. 1992, of Chicago Press, Chicago.
41. Chae, Y.-J., et al., *Generation of co-speech gestures of robot based on morphemic analysis.* Robotics and Autonomous Systems, 2022. **155**: p. 104154.
42. Kang, H. and J. Yang. *Selection of the Optimal Morphological Analyzer for a Korean Word2vec Model*. in *Proceedings of the Korea Information Processing Society Conference*. 2018. Korea Information Processing Society.
43. Igualada, A., N. Esteve-Gibert, and P. Prieto, *Beat gestures improve word recall in 3- to 5-year-old children.* Journal of Experimental Child Psychology, 2017. **156**: p. 99-112.
44. Pérez-Mayos, L., M. Farrús, and J. Adell, *Part-of-speech and prosody-based approaches for robot speech and gesture synchronization.* Journal of intelligent & robotic systems, 2020. **99**(2): p. 277-287.
45. Matsumoto, D. and H.S. Hwang, *Emblematic gestures (emblems).* The Encyclopedia of Cross-Cultural Psychology, 2013. **2**: p. 464-466.
46. Krahmer, E. and M. Swerts, *Audiovisual prosody—introduction to the special issue*. 2009, SAGE Publications Sage UK: London, England. p. 129-133.
47. Zabala, U., et al. *Learning to gesticulate by observation using a deep generative approach*. in *Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings*. 2019. Springer.
48. Camgoz, N.C., et al. *Sign language transformers: Joint end-to-end sign language recognition and translation*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
49. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.
50. Rebol, M., C. Güti, and K. Pietroszek. *Passing a non-verbal turing test: Evaluating gesture animations generated from speech*. in *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. 2021. IEEE.
51. Ronneberger, O., P. Fischer, and T. Brox. *U-net: Convolutional networks for biomedical image segmentation*. in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. 2015. Springer.
52. Chiu, C.-C., L.-P. Morency, and S. Marsella. *Predicting co-verbal gestures: A deep and temporal modeling approach*. in *Intelligent Virtual Agents: 15th International Conference, IVA 2015, Delft, The Netherlands, August 26-28, 2015, Proceedings 15*. 2015. Springer.
53. Cassell, J., H.H. Vilhjálmsson, and T. Bickmore. *Beat: the behavior expression animation toolkit*. in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. 2001.
54. Ferstl, Y., M. Neff, and R. McDonnell. *Multi-objective adversarial gesture generation*. in *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games*. 2019.

55. Latoschik, M.E., et al. *Utilize speech and gestures to realize natural interaction in a virtual environment*. in *IECON'98. Proceedings of the 24th Annual Conference of the IEEE Industrial Electronics Society (Cat. No. 98CH36200)*. 1998. IEEE.
56. Khundam, C. *First person movement control with palm normal and hand gesture interaction in virtual reality*. in *2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. 2015. IEEE.
57. Chen, Y.-L. and F. Lin. *Safety control of discrete event systems using finite state machines with parameters*. in *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148)*. 2001. IEEE.
58. Feng, Z., B. Yan, and T. Xu, *3D direct human-computer interface paradigm based on free hand tracking.* Chinese Journal of Computers, 2014. **37**(6): p. 1309-1323.
59. Lin, W., et al. *Design of hand gestures for manipulating objects in virtual reality*. in *Human-Computer Interaction. User Interface Design, Development and Multimodality: 19th International Conference, HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part I 19*. 2017. Springer.
60. Ferstl, Y., M. Neff, and R. McDonnell, *ExpressGesture: Expressive gesture generation from speech through database matching.* Computer Animation and Virtual Worlds, 2021. **32**(3-4): p. e2016.
61. Lennes, M., et al., *Comparing pitch distributions using Praat and R.* Phonetician, 2015.
62. Kim, H., S.S. Kwak, and M. Kim. *Personality design of sociable robots by control of gesture design factors*. in *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. 2008. IEEE.
63. Kendon, A., *Gesticulation and speech: Two aspects of the process of utterance.* The relationship of verbal and nonverbal communication, 1980. **25**(1980): p. 207-227.
64. Eyben, F., M. Wöllmer, and B. Schuller. *Opensmile: the munich versatile and fast open-source audio feature extractor*. in *Proceedings of the 18th ACM international conference on Multimedia*. 2010.
65. Buehler, P., A. Zisserman, and M. Everingham. *Learning sign language by watching TV (using weakly aligned subtitles)*. in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009. IEEE.
66. Everingham, M., J. Sivic, and A. Zisserman. *Hello! My name is... Buffy"--Automatic Naming of Characters in TV Video*. in *BMVC*. 2006.
67. Lovins, J.B., *Development of a stemming algorithm.* Mech. Transl. Comput. Linguistics, 1968. **11**(1-2): p. 22-31.
68. Eyben, F., M. Wöllmer, and B. Schuller. *OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit*. in *2009 3rd international conference on affective computing and intelligent interaction and workshops*. 2009. IEEE.
69. Bungeroth, J. and H. Ney. *Statistical sign language translation*. in *sign-lang@ LREC 2004*. 2004. European Language Resources Association (ELRA).
70. Koller, O., S. Zargaran, and H. Ney. *Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
71. Koller, O., et al. *Deep sign: Hybrid CNN-HMM for continuous sign language recognition*. in *Proceedings of the British Machine Vision Conference 2016*. 2016.
72. Camgoz. *PHOENIX14T (RWTH-PHOENIX-Weather-2014T)*. Available from: https://paperswithcode.com/dataset/phoenix14t.
73. Starner, T., J. Weaver, and A. Pentland, *Real-time american sign language recognition using desk and wearable computer based video.* IEEE Transactions on pattern analysis and machine intelligence, 1998. **20**(12): p. 1371-1375.

74. Tamura, S. and S. Kawasaki, *Recognition of sign language motion images.* Pattern recognition, 1988. **21**(4): p. 343-353.
75. Rastgoo, R., et al. *Sign language production: A review*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.
76. Kiani, K., S. Rezaeirad, and R. Rastgoo, *HMM-Based Face Recognition Using SVD and Half of the Face Image.* Modeling and Simulation in Electrical and Electronics Engineering, 2021. **1**(2): p. 45-50.
77. Camgöz, N.C., et al., *Rwth-phoenix-weather 2014 t: Parallel corpus of sign language video, gloss and translation.* CVPR, Salt Lake City, UT, 2018. **3**: p. 6.
78. Kipp, M., A. Heloir, and Q. Nguyen. *Sign language avatars: Animation and comprehensibility*. in *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11*. 2011. Springer.
79. Saunders, B., N.C. Camgoz, and R. Bowden. *Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
80. Gibet, S., et al., *Interactive editing in french sign language dedicated to virtual signers: Requirements and challenges.* Universal Access in the Information Society, 2016. **15**: p. 525-539.
81. Guo, D., et al. *Hierarchical LSTM for sign language translation*. in *Proceedings of the AAAI conference on artificial intelligence*. 2018.
82. Stoll, S., et al., *Text2Sign: towards sign language production using neural machine translation and generative adversarial networks.* International Journal of Computer Vision, 2020. **128**(4): p. 891-908.
83. Camgoz, N.C., et al. *Multi-channel transformers for multi-articulatory sign language translation*. in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. 2020. Springer.
84. Zelinka, J. and J. Kanis. *Neural sign language synthesis: Words are our glosses*. in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020.
85. Kusters, A. and S. Sahasrabudhe, *Language ideologies on the difference between gesture and sign.* Language & Communication, 2018. **60**: p. 44-63.
86. Jepson, J., *Urban and rural sign language in India.* Language in Society, 1991. **20**(1): p. 37-57.
87. Cao, Z., et al. *Realtime multi-person 2d pose estimation using part affinity fields*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
88. Klein, B., L. Wolf, and Y. Afek. *A dynamic convolutional layer for short range weather prediction*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.