GESTURE GENERATION FROM URDU TEXT BASED ON DEEP LEARNING APPROACH

By

HUSSAN FATIMA

NATIONAL UNIVERSITY OF MODERN LANGUAGES

ISLAMABAD

February 2024

Gesture Generation from Urdu Text Based on Deep Learning Approach

By

Hussan Fatima

BSCS, National University of Modern Languages, Islamabad, 2020

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In **Computer Science**

To

FACULTY OF ENGINEERING & COMPUTING

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 Computing for acceptance.

Thesis Title: Gesture Generation from Urdu Text Based on Deep Learning Approach

Submitted By: Hussan Fatima **Registration #:** 67 MS/CS/S22

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. Muhammad Farhad 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 (FEC) — Signature of Dean (FEC)

February 10th , 2024

AUTHOR'S DECLARATION

I Hussan Fatima

Daughter of Muhammad Fareed

Registration # 67 MS/CS/S22

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 **Gesture Generation From Urdu Text Based on 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

Hussan Fatima

Name of Candidate

10th February, 2024

Date

ABSTRACT

Title: Gesture Generation from Urdu text based on Deep Learning Approach

When interacting with Artificial Intelligence (AI) machines such as robots and digital assistants, nonverbal cues, including gestures, are crucial alongside speech. Research to endow computers with human-like abilities has been ongoing for years, but most gesture generation studies have been constrained by dependencies on speech input, the English language, and specific speakers. This thesis aims to address these limitations by developing a gesture-generation model that produces high-quality gestures in response to Urdu textual input without requiring speaker assistance.

We have developed our gesture model using Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) algorithms, training it on a custom-created dataset. The implementation results demonstrate the model's effectiveness, achieving a Percentage of Corrected Keypoints (PCK) value of 75% and a mean absolute error rate of 0.3. These outcomes confirm the model's success in generating accurate and reliable gestures for AI interactions in the Urdu language. This research not only tackles the technical challenges of gesture generation but also contributes to the broader goal of enhancing AI's ability to interpret and respond to human nonverbal cues across different languages and cultural contexts.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BLSTM	-	Biconditional Long-Short Term Memory Network
CNN	-	Convolutional Neural Network
CF	-	Communicative Functions
CUDNN	-	CUDA Deep Neural Network
CVAE	-	Conditional Variational Auto-Encoder
DNN	-	Deep Neural Network
GAN	-	Generative Adversarial Network
GRNN	-	Generalized Recurrent Neural Network
GRU	-	Gated Recurrent Unit
HMM	-	Hidden Markov Models
LSTM	-	Long-Short Term Memory Network
MAE	-	Mean Absolute Error
Mres-LSTM	-	Multistream Residual Long-Short Term Memory Network
PCK	-	Percentage of Corrected KeyPoints
RNN	-	Recurrent Neural Network
seMG	-	Surface Electromyography
SSE	-	Sum of Squared Errors
SSIM	-	Structural-Similarity Index
SS-LSTM	-	Sequentially Supervised Long-Short Term Memory Network
SVM	-	Support Vector Machine
TSN	-	Temporal Segment Network
UK	-	United Kingdom
VR	-	Virtual Reality
VRNN	-	Variational Recurrent Neural Network
YOLO	-	You Only Look Once

LIST OF SYMBOLS

α	-	PCK Thershold

ACKNOWLEDGMENT

First, I wish to express my gratitude and deep appreciation to Almighty Allah, who made this study possible and successful. This study would not be accomplished unless the honest espousal that was extended from several sources for which I would like to express my sincere thankfulness and gratitude. Yet, there were significant contributors for my attained success, and I cannot forget their input, especially my research supervisors, Asst. Prof. Dr. Moeenuddin Tariq and Mr. Muhammad Farhad Riaz, who did not leave any stone unturned to guide me during my research journey.

A special thank you to esteemed mentor and my friend Tehmeema Irfan for giving me the best advice on how to get over my challenges. I shall also acknowledge the extended assistance from the administrations of the Department of Computer Sciences who supported me all through my research experience and simplified the challenges I faced. For all whom I did not mention but I shall not neglect their significant contribution, thanks for everything.

DEDICATION

This thesis work is dedicated to my parents and my teachers throughout my education career who have not only loved me unconditionally but whose good examples have taught me to work hard for the things that I aspire to achieve.

CHAPTER 1

INTRODUCTION

1.1 Overview

When two people communicate or talk they do it via two channels, one is the uttered speech or audio by any individual and the other is the movements of the body parts, especially upper body parts. This nonverbal communication channel improves the chance of mutual understanding between two individuals as it accompanies the uttered words. Technology is getting advanced day by day because it reduces human efforts and time, so every individual is expected to perform all the required tasks using advanced technology. Machines are getting human appearance in the form of Robots and Virtual agents [1] as well as in terms of behavior they are more expected to do as human beings.

Artificial Intelligence and its domains including Deep Learning and Machine Learning are playing a crucial role in minimizing the space between Human Beings [2] and Artificial Machines and giving more and more abilities like human beings to machines. Much work has been done to produce human-like behavior for artificial machines but most of the work has certain limitations and constraints where language is the main entity [3]. English is considered as official language so most of the deep learning and machine learning models are built on English language words while other cultural and local languages are never taken under consideration. This research is proposed to mainly focus on the Urdu language and to build a deep learning model that produces artificial robotic gestures [4] from Urdu language text.

1.2 Motivation

Advances in technology have turned the whole world into a global village and intelligent systems are being placed to perform most of the industrial [5], academic, and domestic tasks. People all over the world use their cultural and local languages as their common means of communication. Asian people use to speak approximately 1/3rd of all the languages that are spoken. Urdu is one of the common languages spoken in Asian countries [6], especially in Pakistan. There were a lot of additional Urdu-related activities in the United Kingdom cities where most immigrants had settled. Mushairas were held often (and still are) [7], drawing sizable crowds. Our newly formed Anjuman-e-Tariqa-e-Urdu [8], whose president was the late Raja of Mahmudabad, held regular meetings at the Islamic Cultural Center, which is currently the location of the well-known Regents Park Mosque. Urdu literature piqued the interest of British publishers [9], who were eager to take on such works.

The average literacy rate in Pakistan is about 59.3% [10] which means that half of the population is not educated. The consumption of advanced technology depends on education because the means of communication with advances in technology is in the English Language so as a result almost half of the population [11] fails to get benefit from technology. Artificial Intelligence is automating the whole world. All tasks are now being performed with a single click but unfortunately, uneducated people in certain areas are unable to get benefits from the advances in Artificial Intelligence. This research is to focus on this problem and integrate the cultural Urdu language with Computer Science. Hence this research is designed to propose a deep learning model that generates artificial gestures using Urdu language. Hybrid Deep Learning approaches [12] are promising to produce quality results in most of the deep learning applications including voice recognition, recognition of certain complex patterns, Image processing tasks, and especially Natural Language Processing. The Hybrid Approach [13]is used in this present research to produce gestures using the Urdu Text Dataset which is collected and prepared in this present research.

1.3 Gestures and its Types

A gesture is an instinctive movement made with the upper body parts including hands, arms [14], and shoulders while speaking. The process of continuously mapping these movements into key points of certain Dimensions is known as an artificial gesture.

1.3.1 Beat Gestures

Beat gestures are the kind gestures that have a rhythm in them, and they are close to the pace of the spoken language. Specific words and phrases that need to be highlighted come under the beat gestures. Examples of beat gestures include Repetition, Rhythm and timing, coordination with speech, natural and unconsciousness, [15] emphasis etc.

Figure 1.1: Beat Gestures [15]

1.3.2 Iconic Gestures

Iconic gestures are used when it is required to visually represent any object in visual form. These gestures are more understood universally as they simply give a symbol of what they represent, and they are closer to the speech of any speaker. Characteristics of iconic gestures include Representation, Imitation, Multimodality, and Spontaneity whereas cupping of hands together, Rotation of hand to show the indication of a shape, pointing to a shape, and

drawing an outline of any shape virtually are some examples of iconic gestures [16]as shown in Figure 1.2.

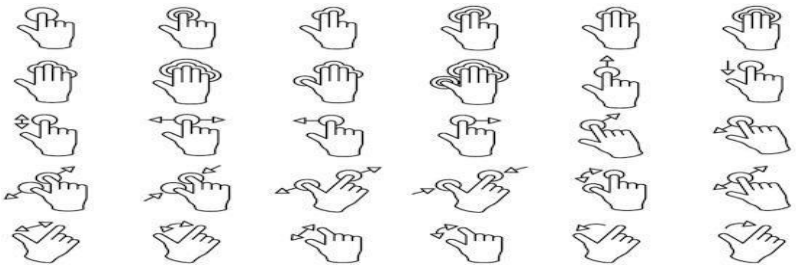

Figure 1.2: Iconic Gestures [16]

1.3.3 Pointing Gestures

Pointing gestures are used to withdraw attention towards any object such as indicating or directing any specific location, when there need to highlight the importance of some kind of object these pointing gestures are used, to make any request or some kind of assistance, showing agreement or disagreement in any aspect [17]as shown in Figure 1.3.

Figure 1.3: Pointing Gestures [17]

1.3.4 Emblematic Gestures

These types of gestures are the special signs that are understood by a community, or a specific group and they commonly include non-verbal symbols and have specific meanings. The meaning of these gestures can be culturally specific, and the use of such gestures can evolve with time and can be changed across generations. Examples of emblematic gestures include Thumbs up, Okay sign, Shaking and nodding of the head, high-five, and peace (V-sign) [18] as shown in Figure 1.4.

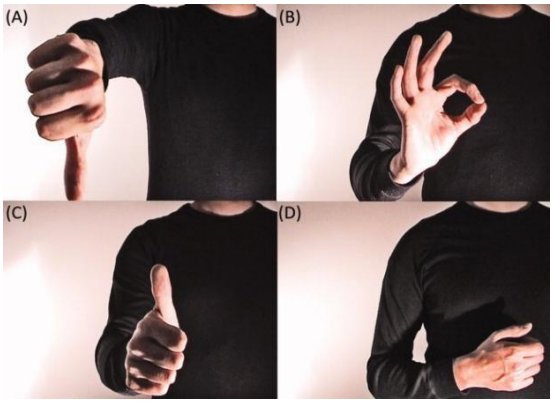

Figure 1.4: Emblematic Gestures [18]

1.3.5 Illustrative Gestures

Illustrative gestures are the kind of gestures that include the movement of body-specific arms and hands. They are usually used to convey ideas or relationships between two things. Characteristics of illustrative gestures are Spatial Representation Enhancing Description, and Cultural Variability. Some common examples of these gestures include Showing the size or length of some shapes, Movement Depiction making outlines, and demonstrating some shapes in the air to show them virtually [19] like in Figure 1.5.

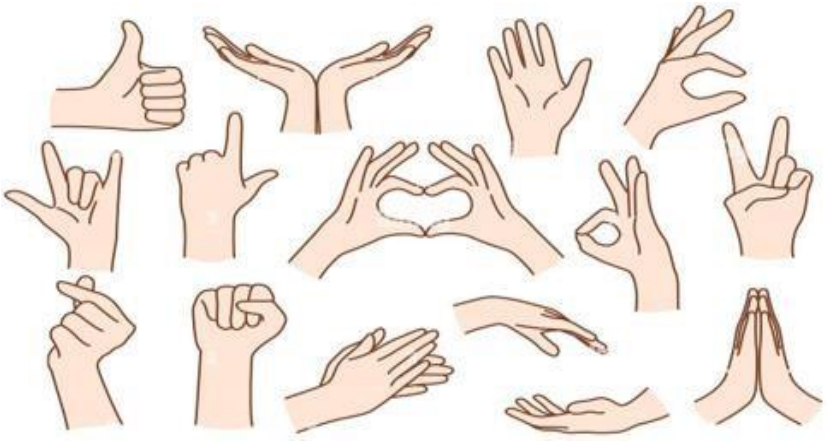

Figure 1.5: Illustrative Gesture [19]

1.3.6 Regulatory Gestures

Regulatory gestures or movements are used to convert or direct the flow of a conversation specifically used to indicate the intentions and to influence the behavior of others during communication. Some common characteristics of regulatory gestures include taking–turn signals, requesting permission, and management of flow, directing and directing attention, controlling the speaking time, and establishing a hierarchy. Some common examples of these gestures are nodding the head, raising and waiving a hand, making a shushing motion [20], etc.

Figure 1.6: Regulatory Gestures [20]

1.3.7 Descriptive Gestures

As clear by its name Description gestures are the gestures that are used while describing something which includes attributes of any object. The key features or common characteristics of descriptive gestures include providing context, clarifying the concepts and the variability, etc. There are some other examples of these gestures such as Demonstrations of motion mimicking any shape,[21] etc. as shown in Figure 1.6.

Figure 1.7: Descriptive Gestures [21]

1.3.8 Adaptors

These are the kind of gestures which are used to fulfill personal needs and undertake several situations. These kinds of gestures are related to discomfort when to show some kind of discomfort these gestures are used. Some characteristics of adaptors include observing behaviors, to show variations that are individual, self-smoothing, and dependent on some context. Some examples of these gestures include playing with hair,[22] biting nails rubbing nails etc

Figure 1.8: Adaptors [22]

1.3.9 Affective Gestures

Affective gestures are the types of gestures that mostly include facial expressions to convey emotions. Characteristics of such motions are facial expressions, body movements, non-verbal communication that include emotions and cultural variability, etc. There are some other examples of such gestures such as raising eyebrows, smiling, seeking of head, [23] etc. as shown in Figure 1.8.

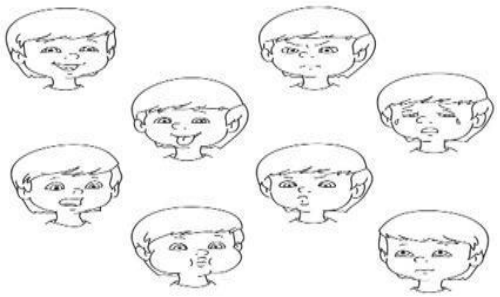

Figure 1.9: Affective Gestures [23]

1.3.10 Emotional Gestures

Emotional gestures are used to convey feelings and emotions which may include behaviors, body movements, etc. The common characteristics of these gestures are body

language, Voice and tone pitch, micro-expressions, etc. Examples of such gestures are tears, laughter, hugging, nodding [24] etc. are shown in Figure 1.9.

Figure 1.10: Emotional Gestures [24]

1.4 Problem Background

Convolutional Neural Network has demonstrated strong performance in text processing and semantic analysis, and hybrid deep learning algorithms have been well-proven in a variety of applications, including picture and text categorization. Consequently, it is possible to generate gestures using hybrid deep-learning techniques. This study's primary goals are to present the hybrid deep-learning gesture model and examine how text input and a sequential model might enhance gesture quality. Using cutting-edge technologies like humanoid robots, which are made specifically to carry out human tasks, has been reducing the barriers to human collaboration in recent years. However, it has some drawbacks and difficulties, like the accuracy of the gestures that are made, which determines the quality. [25,26]. Numerous methods exist for producing gestures from speech, all of which are inspired by the way humans communicate, which combines speech and co-verbal movements [27, 28, 29]. Text is another significant entity that is frequently used by artificial systems, in addition to speech. Local and cultural languages are used as a common language of communication by people all over the world. Approximately one-third of the world's languages are spoken by Asians. One of the common languages spoken in Asian nations, particularly in Pakistan, is Urdu. In the UK cities where most of the immigrants had settled, there were plenty of extra Urdu-related events. Musha'iras drew large crowds and were held frequently (and still do). As an illustration: A consumer can configure a humanoid robot to wait on them and bring them the item they choose from the menu. The customer may use text to guide their decision. The

gesture-generating technique from text input modality has very little literature available [30]. The scope of recently proposed gesture models is limited since they rely on speaker-specific data, which can be in the form of text or audio [31, 32]. based on what is currently known about generalization models. [33]. Thus, designing such generic input-based models is necessary. Since gestures generated by any kind of input require a certain order.

Many sequential models from deep learning, like Long Short-Term Memory and Recurrent Neural Networks, are employed in situations like this when long-term dependencies are necessary. Accuracy and output quality may both rise with the integration of a sequential model.

1.5 Problem Statement

Gestures are crucial in conveying the meaning of words during communication. Previous deep learning and machine learning approaches are well proven in creating gestures, yet they undergo certain limitations like speech-input, language, and dependency of an individual speaker [34]. Addressing these issues, this research thesis aims to produce gestures in response to Urdu textual input. By turning the video data into text, the model creates the gesture based on Urdu language.

1.6 Research Questions

i. How to modify the gesture generation model to use the Urdu Text Input Modality?
ii. How does the Hybrid CNN-LSTM deep learning model impact gesture quality from Urdu text?

1.7 Aim of Research

Designing and developing a modified Deep Learning model to produce human-like gestures for Artificial Intelligent Systems is the goal of this research project. This research's primary goal is to enhance the quality of artificial gestures through the sequential model, which addresses the need for real-time hand or gesture tracking systems, vision, and gesture recognition systems.

1.8 Research Objectives

i. To propose a gesture generation model utilizing Urdu text input modality.
ii. To enhance gesture quality using a hybrid CNN-LSTM deep learning model with Urdu text input.

1.9 Scope of Research Work

The study's focus is restricted to arm, hand, wrist, and shoulder gestures; other body parts are not included in this scope. Furthermore, the only textual input used for gesture recognition is in the English language, which does not take into consideration the variety of languages and [35] non-textual input techniques that are frequently employed in real-world situations. The quality and diversity of the data sets utilized for training and evaluation will determine how effective the research is, and this may provide challenges in fully capturing the range of real-world gesture variability. Furthermore, this study is not intended to address hardware or device-specific issues that may affect the accuracy of gesture detection.

1.10 Thesis Organization

The remaining Thesis is organized as described below:

The research domain and a thorough summary of the body of current literature are presented in Chapter 2. Literature is categorized according to a variety of methods and approaches, including models, algorithms, and specifically the methodology that was employed. In addition, Chapter 2 lists the current limitations and difficulties that offer a foundation for more study and the knowledge gap that results in the creation of a modified text-based gesture model.

In-depth information about the research methodology is provided in Chapter 3, which also offers solutions to address the current constraints and a comprehensive overview of all benchmark methodologies now in use. The approach chosen to carry out this research is described in this chapter. It also provides information on the implementation tools and the evaluation process for the suggested paradigm.

Since Chapter 4 provides a detailed overview of the model's suggested design, it can be regarded as the thesis' central chapter. discussing the tools, models, and algorithms needed to put the suggested architecture for creating gesture models into practice. This chapter discusses specifics of data and whether the requirements for data processing are satisfied. Additionally, this chapter's flowcharts, figures, and diagrams help readers better comprehend the architecture of the model.

The reader will receive an assessment of the suggested model in Chapter 5 along with the parameters that were used to assess the model's recommended architecture. The efficacy of this architecture is further ensured by a comparison of the suggested model with benchmark data sets supplied in this chapter. The accuracy that was attained by comparing the suggested method to different models is shown in a table.

An overview of the research work's contributions and recommendations for further work are provided in Chapter 6.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

The gesture of robots has been analyzed by scholars in past years and this is the topic on which most research is being performed. In this chapter, a brief review of the previous techniques to produce movements is provided. Generating Robotic life-like gestures is difficult i.e. movements synchronization concerning voice. The presented research in this thesis gives a distribution and makes a study available based on a Convolution Neural Network to produce gestures in section 2.2, and sections 2.3 and 2.4, respectively, cover methods pertaining to Morphemic Analysis and Generative Adversarial Networks (GANs). Section 2.5 explains techniques from Data-Driven Approaches. Highlighted Research gaps and directions are discussed in section 2.9. Finally, there is a summary of this chapter in Section 2.10.

2.2 Convolutional Neural Networks

Gesturet popular deep learning method for surface electromyography(seMG)-based gesture detection is CNN architecture, which is split into two groups according to various assessment techniques. Improving intra-session evaluation recognition accuracy is the main goal of the first study [36, 37, 38]. Atzori et al. [39] created spatially and temporally resolved sEMG pictures and trained a CNN algorithm to extract high-level characteristics. To extract spatial information from the instantaneous sEMG pictures, Geng *et al*. [40, 41] developed a unique CNN model and attained state-of-the-art performance. The difference between sessions or subjects is the focus of the second investigation [42, 43]. The CNN model is adapted by Park and Lee [44] to improve the qualities that are learned for inter-subject evaluation. Domain

adaptation based on GengNet [45] was used by Du et al. [46] to increase inter-session accuracy. After Zhai et al. [47] retrieved useful information from the sEMG spectrogram to produce sEMG images, CNN-based architecture was used to simulate the link between gesture labels and sEMG images. Ordo nez and Roggen [48] proposed a deep hybrid CNN-Recurrent Neural Network (RNN) for activity recognition with multimodal wearable sensors. This design provided a natural sensor fusion and described the temporal information of the activity. [49] Wang et al. [50] proposed a novel CNN-Long Short-Term Memory (LSTM) model to solve the pose estimation and gesture detection problems with only RGB videos. The CNN block was utilized to extract the spatial characteristics of each frame, and the sequentially supervised LSTM (SS-LSTM) was proposed to supervise the learning process, which replaced the class label with extra information. Wang et al. used the Hidden Markov Model method to model and reconstruct the dynamic gesture trajectories. An invariant curve moment represents the global feature, and a direction that characterizes the gesture trajectory for recognition represents the local feature. Oreifej et al.[51] used the histogram approach instead of the sequence model to express the spatial and temporal information present in the depth sequence to achieve the identification goal. Chen et al. [52] used the Fourier descriptor method to extract the feature vector, the hidden Markov chain was utilized for recognition, and the hand segmentation approach was utilized to extract the shape and time features from the data set. The biorthogonal wavelet transform was used by Rahman et al. [53] to preprocess the image before constructing a multi-class support vector machine for recognition. These methods have some accuracy already, but not enough to be considered robust. Cheng et al. [52] used convolutional neural networks in conjunction with images of sEMG features to enhance the recognition effect and identify gestures. This successfully addressed the sEMG gesture recognition shortcomings of conventional machine learning. They also combined the extraction of deep abstract features with the 1-dim convolutional kernel. Liao et al. compared the front-end networks and investigated the single multi-box detector (SSD) method. MobileNets was the preferred front-end network, and the MobileNets-SSD network was improved. effectively fixes the problem of hand shading. Li et al. extracted the sEMG signals of the forearm muscles based on hand movements of humans, using three feature values: wavelength, root mean square, and nonlinear feature sample entropy in the time domain. In the end, hand motion recognition with a high accuracy rate was achieved by Generalized Regression Neural Network (GRNN) and Support Vector Machine (SVM). Huang et al. improved the You Only Look Once (YOLO) v3 algorithm based on an empirical criterion to determine whether a worker meets the requirements for wearing a helmet. More noteworthy was the advancement over the initial YOLO v3 algorithm.

Huang et al. developed a framework for semantic segmentation of visual networks with cooperative target detection. By adding parallel semantic segmentation branch operations to the target detection network, a novel multi-vision task combining object categorization, detection, and semantic segmentation is implemented. It effectively improves vision tasks under difficult conditions. Yang et al. proposed a multistream residual network (MResLSTM) for dynamic hand action recognition. The network integrates residual and convolutional short-term memory models into a single, coherent architecture by using a clockwise grouped convolution and channel shuffling technique to reduce the number of network calculations. The result is an extremely accurate identification. Weng et al. developed a cascaded two-level convolutional neural network model and introduced an Angle-Net model to precisely estimate the grasping angle in response to the inaccuracy of previous pose detection techniques. It effectively fixes the problem of multiple objects piled on top of each other, hiding the target and making it difficult for the robot to locate it during a grasp. A weighted adaptive algorithm that integrates multiple factors has been developed by Duan and colleagues to improve RGB-D information processing. Finally, trials are used to verify the resilience and viability of the algorithm. Liu et al. proposed a novel end-to-end dual-stream topology called the fusion of space-time network. This network closely combines spatial and temporal data to acquire rich spatio-temporal information and generate excellent recognition results. Karpathy et al. suggested a multi-resolution CNN network that can manage enormous volumes of data. Its performance has significantly improved as compared to the network that used strong features. Simonyan and associates constructed a CNN model with two streams. The two-channel model consists of a temporal network trained on the optical flow frame and a spatial network trained on the original frame. To employ a segment-based sampling and aggregation module, Wang et al. developed the temporal segment network (TSN), a novel video-based action detection framework. The dual-stream convolutional network served as its model. Model the long-distance time structure. Molchanov et al. combined a high-resolution network (HRN) and a low-resolution network (LRN) to create a new CNN-based categorization network. The recognition result is obtained by probabilistic fusion of the two branches. Deep learning-based gesture recognition is the industry standard in computer vision due to its many scalability and stability advantages. Ginosar et al. generated a dataset of ten speakers' conversational gesture motions as determined by an automated posture identification system. Additionally, they trained models for gesture generation based on convolutional neural networks (CNNs). After being trained with data from a single speaker, the gesture generation model predicts the 2D coordinates of the speaker's hands and arms based on the aural characteristics of speech.

2.3 Recurrent Neural Networks

The RNN architecture has been applied to sEMG-based hand tasks such as posture estimation and sEMG feature extraction. Hioki and Kawasaki [54] developed a neural network with recurrent structure to forecast finger joint angles from sEMG signals. Quivira et al. [55] presented a sEMG-based hand position estimation method using RNN with LSTM cells. With the use of sEMG signals, it developed a model that accurately represented the kinematics of the hand stance and predicted the kinematics of the hand joints. Multiple views: Bootstrapping allows the identification of hand important spots by first creating labels in multiple hand views. These detections are then used in a multi-view geometry to triangulate in three dimensions. "There often exists an unconcluded view, even in cases where there is significant occlusion in a particular image of the hand. Based on this observation, Tomas et al. [56] present a novel technique called multi-view bootstrapping for generating motions with a multi-camera configuration. To improve the detector's performance during training, the newly formed triangulations were used as new labels. The resulting key-point detector offered precision on par with real-time RGB operation when compared to deep sensors. Amor et al. [57] recorded sEMG signals using a Myo armband to identify sign language. They then extracted features from sequential data using an RNN architecture to interpret motions in sign language. The hybrid CNN-RNN architecture has shown promising results in video and wearable sensor recognition. Ebrahimi Kahou et al. [58] introduced a hybrid CNN-RNN architecture for facial expression analysis. The CNN and RNN modules of the architecture were trained independently. Wu et al. [59] developed a hybrid deep learning system that consists of two hybrid CNN-RNN architectures and a regularized fusion layer in order to extract spatial, short-term, and long-term characteristics.

Features were extracted in the presented work in Phinyomark feature set of each channel to generate new sEMG images because the electromyogram signal is noisier than other wearable sensor signals [60]. We then utilize deep neural networks to extract meaningful information between each channel. Nevertheless, compared to regular photos or video frames, the resulting sEMG image is monochrome and has substantially lower pixel resolutions. We add a locally connected layer and meticulously adjust the settings of each layer in our suggested attention-based hybrid CNN-RNN architecture, which has been used for the first time in sEMG-

based gesture detection. A Transformer-based design was put up by Taylor et al. with the express purpose of capturing spatiotemporal patterns in gesture data. Their model demonstrated the promise of transformer-based techniques in this domain by achieving state-of-the-art outcomes in terms of both naturalness and diversity of generated gestures. The Attention-based LSTM architecture, which Cao et al. introduced, is a variation of the LSTM architecture that includes methods to selectively attend to distinct parts of input sequences. This enables the model to focus on relevant context during gesture production. There is potential for this attention process to enhance the coherence and fluidity of created motions. The application of Long Short-Term Memory (LSTM) networks for gesture generation was first described by Graves et al. Their model produced realistic and coordinated hand motions, demonstrating an impressive ability to capture temporal dependencies. The foundation for further research on using RNNs for gesture synthesis was established by this work. To increase gesture synthesis performance even with a limited amount of annotated data, Smith et al. proposed transfer learning techniques that make use of pre-trained models on larger datasets. This addresses the problem of data scarcity. To provide more realistic and expressive gesture generation, Liu et al. concentrated on creating multi-modal gestures by creating a hierarchical RNN architecture that could capture dependencies across various body parts and modalities. In their 2019 study, Zhang et al. examined the application of variational recurrent neural networks (VRNNs) to produce gestures, utilizing latent variables to capture the variability and underlying structure of gesture sequences. Their method made it easier to create a wide variety of realistic movements and provided fresh perspectives on probabilistic modeling methods for gesture synthesis. Jiang et al. addressed concerns about cultural sensitivity in gesture synthesis by examining the creation of gestures that are appropriate for different cultures. Their research demonstrated how crucial cultural context is for gesture generation applications. By including a self-attention mechanism into the LSTM design for gesture production, Park et al. contributed to the field. Their model produced more logical and contextually appropriate gesture synthesis by learning to dynamically balance the significance of various input items.

2.4 Generative Adversarial Networks

The generator takes samples from a noise distribution (z) to produce data and samples. Samples produced by the generator are referred to as fake samples (labeled as 0), and samples derived from the original data are referred to as real samples (labeled as 1). In a minimax game between the generator and the discriminator, the generator's goal is to trick the discriminator by generating samples that are almost identical to the original, while the discriminator's job is to identify actual samples from fake ones. Deep neural networks (DNNs) have been used in several recent research to study speech-driven head motions and positions.[61] Using solely completely connected layers, only bidirectional long-short term memory (BLSTM) units, and a hybrid technique that combines fully linked layers and BLSTMs, Ding et al. [62] investigated DNNs. These models, which were designed to reduce the sum of squared errors (SSE) between the predictions and the original head movements, mapped speech filter bank properties to head movements. Their findings showed that employing the BLSTM model alone outperformed using the fully connected DNN alone. The combination strategy helped them perform at their peak.[63] Greenwood et al. [64] mapped the filter bank features taken from speech to head postures, creating distinct deep models for the speaking and listening turns. They facilitated the comparison of a conditional variational auto-encoder (CVAE) and a BLSTM model by presenting the statistical characteristics of the moments for the produced head movements. The issue of predicting human mobility is connected to the recognition of human gestures and activities. In motion forecasting, the aim is to predict a sequence of future poses conditioned on the input, given a sequence of human skeleton poses. Because human motion is stochastic and has non-linear dynamics, this is a non-trivial challenge. After deep learning techniques demonstrated remarkable results across a broad range of computer vision challenges Martinez et al. [65] used the Human 3.6M dataset to train a sequence-to-sequence Gated Recurrent Unit (GRU) network and obtain state-of-the-art short-term motion prediction results.[66] By using an encoder-decoder architecture, Butepage et al. [67] were able to learn a reliable feature representation of the human skeletal data and predict future 3D poses. In early sEMG-based gesture recognition frameworks, a great deal of research has been done on the handcrafted features and conventional machine learning classifiers. Three types of hand-crafted sEMG-based features now in use are time domain, frequency domain, and time-frequency domain features [68]. Numerous researchers concentrated on either presenting novel sEMG features based on their expertise in the field or examining current features to suggest novel feature sets.

Li et al. proposed a unique GAN architecture designed for challenges involving gesture generation. Their program learned to generate believable and varied gestures that closely mirrored human movements by redefining the generation process as a game between a generator and a discriminator. The groundwork for later studies on the application of GANs to gesture synthesis was established by this work. A conditional GAN (cGAN) framework was presented by Zhang et al. to facilitate the creation of gestures that are contingent on features or circumstances. Their methodology acquired more control over the generated movements by supplying extra information, including position labels or semantic descriptors, which made applications requiring customized gesture synthesis easier to implement. A Progressive GAN architecture was developed especially for synthesizing high-resolution gesture sequences with realistic motion dynamics and fine-grained details by Tulyakov et al. When it came to the temporal coherence and visual clarity of the generated gestures, their model fared better than earlier methods. To increase the diversity of synthesized gestures, Yoo et al. implemented approaches for data augmentation and style transfer. This allowed the model to generate many gestures even with a small amount of training data. Kwon et al. used the GAN framework with cross-modal constraints to study the creation of multi-modal gestures. Their model learned to produce synchronized multi-modal outputs by modeling gestures in conjunction with accompanying textual or audio cues. This made the model useful for applications in multimedia content creation and gesture-based communication systems. A brand-new adversarial imitation learning system for gesture generation was presented by Chen et al. Their model was able to replicate the gates and facial expressions of human demonstrations, producing a highly accurate synthesis of a wide range of realistic gestures. By incorporating a novel self-attention mechanism into the GAN architecture for gesture creation, Park et al. contributed to the field. Their model's capacity to dynamically weigh the significance of different spatial and temporal characteristics led to a more coherent and contextually appropriate synthesis of gestures. Jeremy Chu et al suggested a conditional GAN called Word Gesture-GAN. It accepts any text as input and produces realistic word-gesture motions in two dimensions: temporal (timestamps of touch points) and spatial (touch point coordinates). To provide for control over the variation in generated gestures, it uses a Variational Auto-Encoder to extract and incorporate variations of user-drawn gestures into a Gaussian distribution. The model is useful for developing and accessing gestural input systems since it performs better than other gesture production models currently in use. Minho Le et al. proposed a system that combines a GAN and an autoencoder to generate a series of sequential human behaviors that are conditioned on beginning states and class labels. The autoencoder and GAN are cooperatively optimized during the end-to-end

training procedure. Even though it isn't particularly gesture-focused, it shows how effective GANs are at producing sequential movements. Kalyan Chatterjee et al. proposed a hand gesture classification using a Convolutional Neural Network (CNNs) coupled with Generative Adversarial Network (GANs). This work uses GANs in conjunction with Convolutional Neural Networks (CNNs) to produce a variety of hand motion sets, however it is not solely GAN-focused. Enhancing usability for marginalized communities is the aim, with a focus on hand gesture recognition's importance.

2.5 Database- Driven Approach

Combining different rules may cause contradictory expressions to be synthesized [69]. Scholars have recently focused on the impact of communicative functions (CF) on the inference of animation. A system that is intended to choose motions from recorded multimodal nonverbal behaviors in accordance with CF is described in the work. Additional research focuses on examining how CF contributes to head animation computation and gesture performance generation [70]. Human audiovisual data becomes available with the introduction of motion capture technologies and video-based tracking devices. Many academics use statistical frameworks to extract the intricate temporal relationships from the available human data to create realistic and organic animations. Speech-driven and text-driven statistical frameworks are two categories into which existing data-driven systems can be divided based on the input signals; these categories will be discussed in the following sections. Albert Dipanda, Cyrille Mignot, and Ammar Ahmad, provide an overview of hand object modeling, along with a range of applications to help with the hand gesture problem. They have been greatly influenced by advancements in monitoring, particularly in the detection of the entire human body, as there are similarities in the assessment of the human body and hand. Due to the occurrence of manual poses and the difficulties in rendering and manipulating hand objects, the exacerbated weight of amplitude generated by increased power, the connectional specification on space of hand objects, results in self-occlusion. A framework for early gesture recognition was proposed by Rohit Agrawal *et al.* [19]. A sequence-to-sequence motion forecasting model was trained with a partially observed gesture represented by a set of poses, and it generated a sequence of anticipated poses. The partially observed ground truth gesture and the predicted pose sequence were combined and fed into a random forest gesture categorization algorithm. It demonstrates

that the sequence-to-sequence model's output was added to the partially seen gesture to greatly increase recognition accuracy. Gesture classification accuracy rose from 45% to 87% in studies using the MSRC-12 gesture recognition dataset when a partially observed gesture of 50 frames was enhanced with an extra 25 frames of anticipated motion, and to 93% when augmented with 100 frames of predicted motion. There are some statistical and machine learning approaches. The latent link between speech and gesture in statistical systems is modeled by the statistics of the underlying gesture distribution. Statistical techniques assume less about the speech-gesture link than rule-based systems do. Instead, they either apply a prior probability distribution or precompute conditional probabilities for the gesture data. One of the first statistical systems was put forth by Kipp, who created a gesture profile by evaluating an annotated co-speech dataset and modeling a person's gestures. Using the video annotation application ANVIL, the data was tagged to create a gesture profile with distinct attributes including handedness, timing, and communicative function. After that, the gesture profiles were developed using statistical models that drew inspiration from research on dialogue act and speech recognition. Conditional probabilities on gesture bi-grams and the occurrence of the gesture given semantics from input text were used to determine the plausibility of a gesture. The product was statistical models that formed a person's gesture profile based on their unique handedness, timing, and transitions. After that, realistic motions were produced from the annotated input speech using the profiles. There were several phases to the generating process: 1) Assigning semantic tags to input text; 2) Generating all possible gestures, adding them to an intermediate graph representation, and labeling the graph with probability estimates; 3) Using text-gesture associations and timing profiles, respectively, to filter and temporally arrange gestures. An XML action script that may be utilized in a system for downstream animation was the outcome in the end. Neff et al. expanded on this strategy by presenting a statistical system that not only included a character-specific animation lexicon but also learnt gesture profiles. The system was divided into two stages. A hand-annotated video corpus of a character in ANVIL served as the basis for the pre-processing phase. Like Kipp's annotation procedure, but with an extra English-speaking character. Based on the annotated data, an animation lexicon and gesture profile (a statistical model) were developed. The latter included information about hand orientation, torso posture, and after strokes (i.e., repeated hand movements that follow a notable stroke) for every gesture lexeme. There were two different routes in the fully automated generation phase: 1) re-creation, which was helpful for verifying the annotations, was able to recreate the gestures (shown in the video) in the animation system after receiving an annotated video as input; 2) gesture creation that might produce gestures without the need for visual input by using newly annotated text. To

create a gesture script, either path made use of the character's gesture profile. Bergman et al. and Kopp et al. proposed a statistical approach for the conversion of speech that describes objects into gestures. By combining relative and imagistic representations of knowledge for content planning and concrete speech and gesture formulation, the proposed method produced synchronized speech and gesture. Using virtual reality (VR), the researchers conducted dyadic talks in which one speaker provides spatial directions to another. The purpose of the study was to determine which contextual elements influence how speech and gesture are formed to describe tangible items. They created a Bayesian network for gesture formulation as part of their methodology. A probability distribution over gesture attributes including indexing, positioning, shaping, drawing, and posing was established by the Bayesian network. The idiosyncratic patterns for mapping visuospatial referents onto gesture morphology—that is, the unique ways in which people might index, shape, or draw gestures while describing referent objects—were also taken into consideration by the probability distribution. Fine-grained features such as hand shape, wrist location, palm direction, extended finger direction, movement trajectory, and direction were produced using gesture formulation. The framework used the rule-based Articulated Communicator Engine to achieve synchronized speech and gesture for the final animation. Levine et al. developed a hidden Markov model to choose the best motion clip from a motion capture database. To choose the gesture sub-units from the motion capture that would best fit the current utterance's tone and ensure a smooth transition, the trained HMM looked for prosody cues. But because prosody and gesture sub-units were closely linked, the system became dependent on the quantity and quality of training data, which led to overfitting. Levine et al. proposed "gesture controllers" that separated the kinematic characteristics of gestures—such as speed and spatial extent—from their shape, making it an improvement over the prior system. Using a conditional random field (CRF) that examined the audio characteristics in the input speech and discovered a distribution over a range of hidden states, gesture controllers deduced the kinematics of gestures. By encoding the latent structure of gesture kinematics without considering the gesture's morphology, the hidden states lessened overfitting by lowering the number of erroneous correlations. Ultimately, a Markov Decision Process (MDP) selected the relevant gesture clips by using an optimal policy that it had learnt using the reinforcement learning algorithm. It did this by taking the hidden states and their distribution as input. Chiu et al. maintained the use of features to train a probabilistic model for gesture production. They limited the scope of their research to learning prosodic gestures, rhythmic movements, or beats. A modified Hierarchical Factored Conditional Restricted Boltzmann Machine served as the foundation for the gesture generator. Using an unsupervised

learning approach, they trained a conditional Restricted Boltzmann Machine to create a compact motion representation initially. The gesture representation was then generated for each time step until the entire motion sequence was finished by the HFCRBM generator using an autoregressive process that took in the preceding gesture representation and a series of audio features taken from the original speech. Lastly, they lessened wrist joint acceleration when it was beyond a certain threshold, which helped to smooth out discontinuities between frames. But because their methodology was limited to rhythmic gestures, it ignored other frequently occurring gesture categories such pantomimes, iconic, deictic, symbolic, and metaphoric gestures. Hasegawa et al. proposed a bi-directional LSTM in an autoregressive manner to produce gestures from auditory utterances. Over a considerable amount of time, the bi-directional LSTM learnt audio-gesture connections with both forward and backward consistencies. Using a headset and marker-based motion capture, a novel audio gesture dataset was used to train the model. At each LSTM timestep, the model used the input of speech attributes to predict a complete skeletal human position. Then, to smooth out discontinuities in the resulting pose sequences, temporal filtering was applied. Methods that employed audio as the main modality resulted in precisely timed hand movements that were mostly in line with beat gestures and were frequently strongly connected with acoustics. The absence of a text transcript, however, means that they were not aware of the context and structural elements of the text, such as punctuation and semantic meaning. A system like this can lead to more expressive and meaningful motions. Thus, we will now go over a few strategies that employed text as the main input modality. A text-based gesture generating technique for operating a humanoid robot was presented by Ishi et al. By linking words to concepts, concepts to gesture categories (such as iconic, metaphoric, deictic, beat, emblem, and adaptor), and gesture categories to gesture motions, they modeled the translation of text to gesture motion. They also generated conditional probabilities to simulate the relationship between gesture categories and motion clusters that were precomputed using the k-means clustering algorithm, as well as the relationship between word concepts and gesture categories. Bhattacharya et al. used text transcripts to generate expressive emotive gestures for virtual agents in storytelling and dialogue scenarios. The actors in MPI-EBEDB, a dataset of multiple emotion categories (sadness, relief, amusement, anger, disgust, fear, joy, neutral, pride, sadness, shame, and surprise), performed these emotions. They used transformer-based encoders and decoders in their method. where the text transcript phrases (encoded as GloVe embeddings) were sent into the encoder, which created an encoding and concatenated it with the agent attributes (gender, handedness, desired emotion, and narration/conversation). The Transformer decoder was used

to produce the joint positions for the subsequent pose by feeding it the encoded concatenation and 3D joint locations from the previous stance. The procedure was repeatedly carried out until the entire pose sequence was produced. Systems for creating gestures that are text-based or audio-based offer an intriguing trade-off. Although they lack semantic context, audio-based generators can produce rhythmic or kinematic gestures (like beats) thanks to their access to prosody and intonation. On the other hand, text-based generators lack prosodic and intonation information, but they do have access to semantic context, which aids in the generation of gestures that convey meaning (such as iconic or metaphoric). Consequently, a gesture generator can learn to generate semantically appropriate and rhythmic co-speech gestures by fusing the textual and auditory modalities. While producing meaning-carrying gestures solely through auditory means is theoretically feasible, it is improbable since prosody is appropriate for kinematics but insufficient to deduce shape, which is linked to meaning [LWH 12]. To the best of our knowledge, meaningful gestures from spoken audio alone have not been proven through empirical research. Rather, it seems that the most promising method for producing meaningful gestures to yet is the combination of text and music. As a result, we concentrate on methods that integrate these two modalities to produce expressive gestures that convey meaning. A method that produced co-verbal gestures by fusing the speech prosody and text was presented by Chiu et al. A fully connected network was used for representation learning in their model, which they named the Deep Conditional Neural Field, and a Conditional Random Field was used for temporal modeling. To predict a series of gesture signs—a collection of predetermined hand motions—the model used prosody features, part-of-speech tags, and a text transcript as input for the gesture prediction task. The next natural step was to use the representation power of deep learning models for multimodal input (text and audio) to generate co-speech gestures. Yoon et al., Ahuja et al., and Kucherenko et al. were the three distinct research groups that suggested the first deep-learning-based gesture generators that generated continuous motions using both audio and text. Next, we talk about their groundbreaking work fusing text and audio, and then later initiatives in the field. Given that co-verbal gestures are impromptu, incredibly unique, and nonperiodic, animating them remains a very difficult task. Rule-based methods use motion recording to produce well-formed gestures, however they are rigid and don't provide a variety of motions.

2.6 Sign Language Processing (SLP)

Rule-based and statistical-based methods can be used to categorize machine learning-based algorithms. Iwai et al. used a glove-based machine learning system, one of the statistically based techniques, to categorize hand motions [32]. They first used the nearest-neighbor method to extract features, and then they used a decision tree algorithm to classify the data. Wilson and colleagues used a statistically based concealed. They obtained good results with the RGB hand motion dataset by concentrating primarily on the max-pooling layer of the CNN [37]. The primary issue with this research is the difficulty in identifying hand gestures that include orientation fluctuations and partial occlusions. Tao et al. used a CNN model with multi-view augmentation to identify kinetic-sensor-based hand motions for American Sign Language (ASL) to get over this difficulty [38]. Facial expression manipulation has been studied with audio-based motion generation [5], and gesture production has been studied with a similar methodology [17]. An objective function is defined as a model.

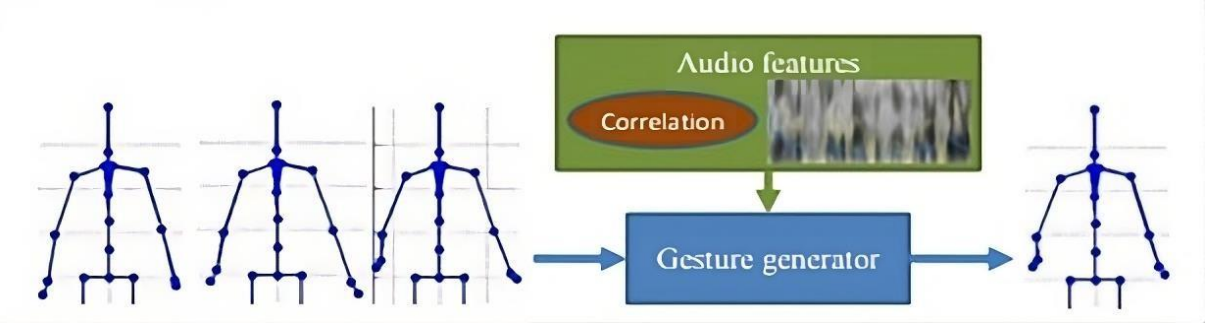

Figure 2.1: The Architecture of the Gesture Generation Process [17]

The relationship between utterances and motions as well as the sequential relationship among motions. A series of animation segments are then collected, and gesture animation is created by optimizing the objective function. This approach is like prior efforts. This architecture is compatible with Hidden Markov models (HMM), which have been used to produce arm gestures [17] and head motions [6, 19]. Two types of features, such as hand-crafted and deep-learned features, have also been studied for skeleton-based action recognition.

The 3D joint positions of the human body were utilized in manually constructed feature-based methods to calculate skeletal quad [8], points in a Lie group [29], Action let Ensemble [32], and Eigen Joints [35]. Dan Huang [30] presented a variation adversarial autoencoder-based approach for 6D posture estimation that is limited to using RGB data.

This network's encoder and decoder have structural symmetry, and the encoder uses a hierarchical stream-by-stream approach to extract features. The orientation feature is implicitly represented using the latent space that was obtained from the RGB image. It is practically possible to realize the 6D posture estimation of the item based on the template matching method. Several CNN- and Recurrent Neural Network-based architectures are used in deep learning-based techniques [6], Dan Gelb, Henrique Weber, and Claudio Rosito-Jung (2016) [5], Research indicates that a new wave of processing and display technologies has sparked the development of user interfaces that transform commonplace objects, such as tables and walls, into immersive planes. When the hand is laid flat, the color information and depth may effectively divide the hand's shape at different distances from the surface. This is part of the goal of making computing resources available, adaptable, and allocated.

The spoken words were interpreted by the encoder, and human emotions were produced by the decoder. The model created frame-by-frame poses of gestures against a natural language without requiring any prior information. The resulting 2-D poses were then transformed into 3-D by using a Robot Prototype. Co-speech gestures have been studied by many psychologists. This large body of research has mostly relied on examining a small number of chosen subjects using synchronized story retellings recorded in laboratory settings. By adjusting the noise in the hand skeleton data, they concentrated on the noisy dataset and, for the 14 and 28 gestures in the DHG dataset, respectively, obtained accuracy of 80.44% and 85.92%. To identify a temporal 3D position, Nunez et al. suggested combining CNN and LSTM models. They reported 85.46% and 81.10% accuracy for the 14 and 28 gestures in the DHG dataset, respectively [17].

To generate node and edge properties for spatial and temporal domains, we first used a deep neural network and subsequently a spatial-temporal and temporal-spatial branch. The analysis for these investigations was done by hand. Rather, we want to evaluate natural conversational gestures through a data-driven approach. [17] In order to cut down on noise, Ma et al. used an unscented kalman filter in conjunction with an LSTM for classification [25]. Wep extracted three broad deep learning features and concatenated them to create the final feature vector to improve the system's generalization. We applied a spatial-temporal mask to lower the computational cost, and for the MSRA dataset, we achieved 94.12% accuracy; for the DHG dataset, we reached 92.00% and 88.78% accuracy. Similarly, using the SHREC'17 dataset, they obtained 97.01% and 92.78% accuracy for the 14 and 28 gestures, respectively.

2.6 Comparisons of Deep Learning SLP models-based Schemes.

Table 2.1 Summary of Deep Learning SLP Models

References	Author	Techniques	Results	Limitations
[33]	Dan Huang, Hyemin Ahn, Shile Li, Yueming Hu and Dongheui Lee	Adversial Autoencoders, Self-Supervised Learning	Mean=38.56	Only use offline detection of 6D pose.
[32]	Abu Saleh Musa Miah, Md. Al Mehedi Hasan, Jungpil Shin, Yuichi Okuyama and Yoichi Tomioka	Spatial Attention Model, Feature extraction using CNN	99% accuracy	

[78]	ABU SALEH MUSA MIAH, MD. AL MEHEDI HASAN, and JUNGPIL SHIN	Multi-branch attention-based graph and a general deep learning model to recognize hand gestures by extracting all possible types of skeleton-based features.	Model achieved 94.12%, 92.00%, and 97.01% accuracy	3D hand skeleton information from gestures to develop a sign language-based communication system.
[80]	Shichen Zhang, Tianlei Wang and Jiuwen Cao	Auto encoder	MMRAEs not only improve the overall accuracy, but also effectively reduce the network size.	Not focus on constrained modeling based MMRAEs 172 by exploiting the feature correlation within the same class as well as cross-classes.
[48]	Ikhsanul Habibie, Michael Nef, and Christian Theobalt	audio-gesture clips from a database using a KNN algorithm and GAN model	This approach outperforms the state of the-art both in terms of naturalness and audio-synchronicity	Limitation of search-based algorithm is the potentially expensive computation time compared to single-pass inference approaches of the purely learning-based counterparts.
[81]	Mireille Fares, Catherine Pelachaud and Nicolas Obin	CNN and transformer decoder	RMSE errors are much smaller than LSTM-based baseline model of prediction. Model received similar values for the 5 factors.	

[82]	Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja	PHOENIX 14T	Training of the model was done via skeleton annotations.	High Complexity.
[83]		PHOENIX 14T	Gloss information was not used.	High Complexity.
[84]		Czech news	Give better. Results even Skeleton parts are missing.	

2.8 Research Gap and Direction

Several issues were found during a thorough review of the literature on gesture-generating models, which prompted the creation of the text-to-gesture generation model that will be covered in the upcoming chapters. The intricacy of the model, the scope of the input modalities, and the caliber of the gestures that are produced as output are the current problems.

Here is a list of several issues and holes that have been identified:

- Since the field of artificial intelligence is centered on human-machine interaction, there is less and less room between virtual agents and robots, particularly humanoid robots. Thus far, speech has been the basis for this exchange. There hasn't been much research

done on the text, which must be investigated because it's a valuable entity that forms the basis of this interaction.

- Since gestures are a simple means of communication between robots and artificial intelligence, several models regarding deep learning and machine learning have been put out to create artificial gestures. However, these models fall short of human-like motions in terms of quality. 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 the attention of researchers.
- Recurrent Neural Networks (RNN) is the most popular algorithm in the domain of Artificial Intelligence and Deep learning, and it has been proved well to show good performance when processing sequential data. Sequencing the motions is the key to using individual gestures to produce motion. Consequently, research into the potential effects of a sequential model on the sequence and quality of generated gestures is necessary.
- Standard English language keywords are the only ones included in the gesture generating models that are currently available, whether based on text input or voice. Even if most people can understand English, additional languages are still required as the foundation for artificial intelligence models.

Several proposed Models have been developed to generate gestures. These models and techniques have certain constraints and limitations. As discussed in section 2.3 the S2G model proposed in [2] acquires correctly predicted Key-Points using a Convolutional Neural Network (CNN). This research is being conducted to identify a few of the above-mentioned gaps. There is hardly any work found in the literature survey which is conducted that produces gestures other than the English language Considering the Urdu-speaking culture and viewing the importance of other languages than English this work is directed to propose a Text-to Gesture generation model that produces gestures corresponding to Urdu text.

2.8 Summary

This chapter offers a thorough literature review of gestures and the models that are currently in use, which establishes a strong research base. This chapter offers a critical analysis

of the models that are now in use about various methodologies. Additionally, a theoretical discussion of each technique's strengths and weaknesses is provided. Lastly, their shortcomings and restrictions are addressed, which ought to serve as a challenge for more study.

CHAPTER 3

METHODOLOGY

3.1 Overview

This chapter discusses details of the research approach that was used to create the text-to-Gesture model and examines the architecture of the suggested deep learning model. This chapter carefully examines the several stages that make up the research technique. The difficulties encountered throughout the research process are explained, offering an understanding of the study's endeavor. This chapter provides a thorough review of the procedures taken in the research process as well as an in-depth discussion of the design phase. The principal aim of the framework that has been suggested is to produce expressive gestures from Urdu text and evaluate its effectiveness by measuring the percentage of Corrected Key points (PCK). A crucial component is the creation of our own dataset, which is designed to produce improved and notable features for model training, together with the use of the best techniques for hyperparameter tuning. In this chapter, the necessary conditions are covered in detail, providing insight into the data pretreatment methods used to reduce data complexity to facilitate feature extraction and feature vector embedding. This chapter also briefly discusses the size, scope, and instances that make up the dataset.

3.2 Research Methodology

This section elaborates the methodology and organizational framework utilized in the suggested study project. It discusses the step-by-step processes used to reduce the deviation from the objectives. Three distinct phases of the investigation are depicted in Fig. 3.1

The first stage, called the Analysis Phase, entails a thorough review of current plans and various approaches to gesture generation. This includes the fields of morphological analysis, neural networks,[71] generative adversarial networks (GANs), and several database-driven approaches. Neural Networks are particularly well-suited to managing intricate calculations with little data by using parallel processing. Superior motions are produced by Morphemic Analysis, which synchronizes with spoken or textual utterances by understanding words at their origin. Gesture generation is a strong suit for GANs-based techniques, which take advantage of their dual neural network architecture to process information in parallel. Database-driven techniques are useful for analyzing complex data and producing several results. The literature proposes several data-driven methodologies for gesture generation. We find gaps and limits in the literature by carefully examining it; one such drawback is the lack of gesture production from Urdu text. The problem statement and the objectives are shaped by these limitations.

The next stage, which is called the Design and Development Phase, is the most important one since it involves building models and architecture while considering all the necessary resources, such as datasets and model parameters. The model is tuned to take textual input because it recognizes that textual information can yield superior motions. A sequential approach is used to improve gesture output because movements and text sentences—which are shown as word vectors in a sequential order—are synchronous. Convolutional neural networks (CNNs) are used to extract significant features [72] from the input text. After then, a sequential long short-term memory (LSTM) is fed these features to generate two-dimensional important points of gestures. To ensure robustness and dependability, the text-to-gesture model is put through a rigorous testing process that includes multiple epochs and train-test splits.

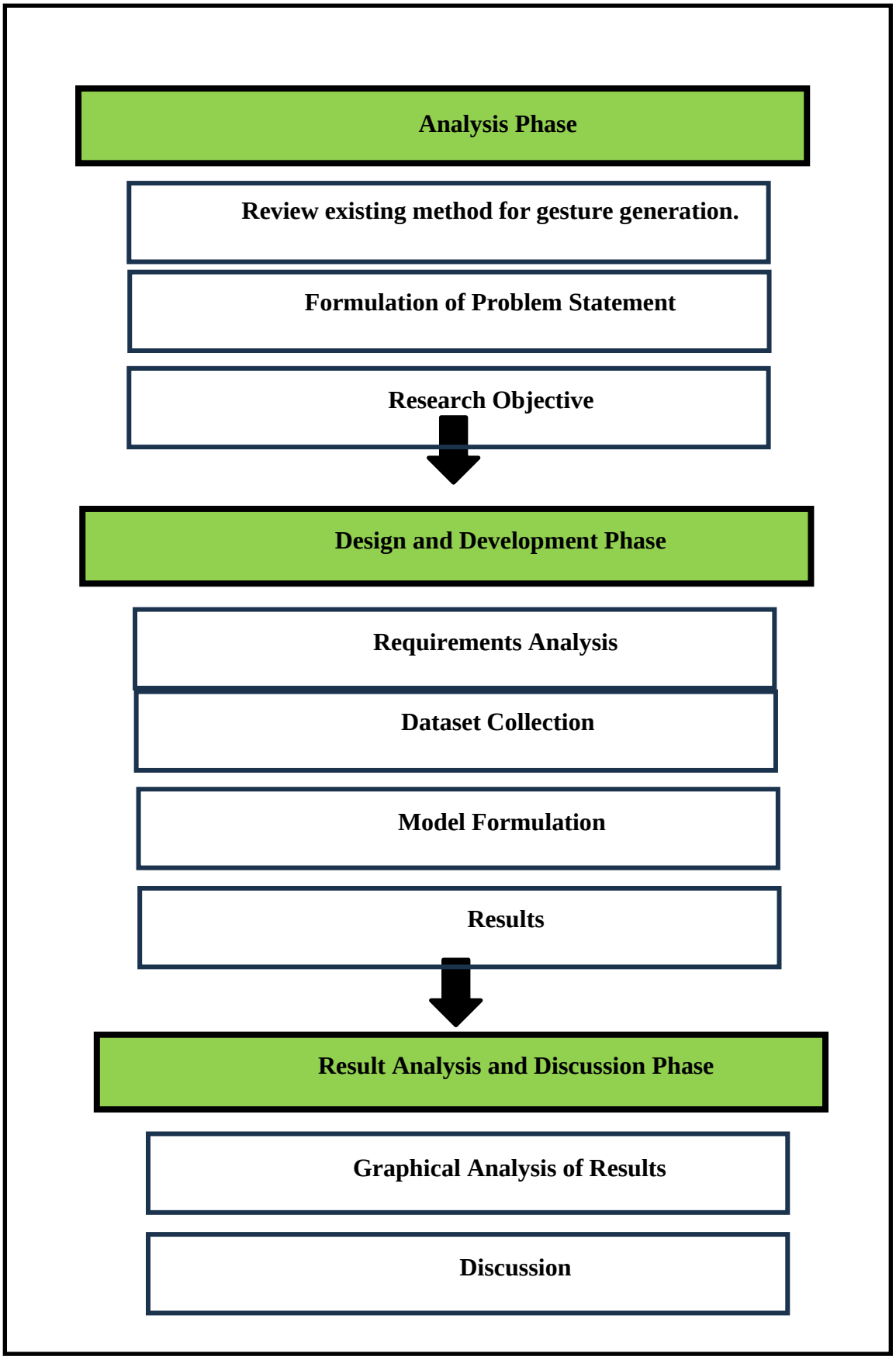

Figure 3.1: Operational Framework of the Research

3.3 Requirement Analysis

Any suggested proposal must first undergo a thorough analysis of needs, which entails a thorough investigation and gathering of all relevant resources. This stage consists of several smaller processes, such as requirements collection, prioritization, documentation, validation, and general management. To put the intended text-to-gesture model into practice, the most important prerequisite is a large dataset in addition to a system that has GPUs and a few other essential requirements for the experiment. In Section 3.2.2, these needs are explained in more detail.

3.3.1 Dataset

This section outlines how the dataset was created. It consists of 15 videos, each lasting 30 minutes. This video records an Urdu professor's speech, with the camera staying in one place the entire time. The Open Pose identification technology made it easier to extract 2D essential points from the videos. Thirty key points representing the articulation of the arms, hands, elbows, wrists, and shoulders were identified. In contrast to other datasets, our collection highlights the linguistic subtleties in the Urdu language, providing a unique basis for our text-to-gesture model. The construction of this dataset is essential to adjusting our research to the unique nuances of Urdu communication patterns, which are in perfect alignment with the research objectives. The dataset was created by distilling the content of 15 videos with a single speaker. The output is a thorough compilation that includes speaker frames, important points, and related language expressions. We arranged 23 different CSV files, each containing an average of 4,000 occurrences, within each word dataset. Two columns make up the structure of these files: one column contains the extracted word, and the other column has the matching time. When no word was found for a given frame, the keyword "BLANK" was assigned to provide clarification. The entire repertory consists of 49 unique key points that cover both x and y coordinates in two dimensions.

3.3.2 System Requirements

System requirements outline the necessary capabilities that a system must have in order to run a certain piece of software. These parameters usually include storage capacity, operating systems, dependencies, CPU specifications, and graphics card concerns. Several crucial conditions have to be met in order for the suggested system [73] used in this study to be put into practice, as listed below:

i. **Python 2.7:** The foundation for executing the suggested model in this study is provided by the notable Python release of 2010, [74] namely version 2.7.
ii. **Cuda 9.0:** With the Compute Unified Device Architecture (Cuda 9.0), [75] programmers may leverage NVIDIA GPUs' computing capacity for a variety of applications, including deep learning.
iii. **CuDNN:** Cuda Deep Neural Network Library (CuDNN): CuDNN makes it easier to do complex deep learning operations, such as pooling, normalisation, and applying different activation functions. It is interoperable with frameworks [76] like PyTorch, Caffe, and NXNet and speeds up these operations on GPUs.
iv. **Open CV:** The Open-Source Computer Vision (OpenCV) library is essential [77] for managing image processing tasks, especially when using gesture frames for training.

3.4 Pre-Processing

'Pre-processing' refers to a set of procedures, various approaches, and methods that are used on unprocessed data to prepare it for use according to certain specifications. Computational interventions are necessary to ensure the quality and effectiveness of data for analytical purposes. Pre-processing, which includes tasks like data normalization and formatting to meet a deep learning model's training criteria, is a crucial and essential stage in the field of deep learning. This aspect is essential to improving the model's accuracy.

The ground truth for the model's training in this study, which focused on general text inputs, was gestures rather than speaker information. 30 two-dimensional (2D) key points representing the arms, hands, wrists, elbows, and shoulders made up the dataset, making the equivalent 60 points in total. It became necessary to eliminate speaker-specific data from words and gestures in the dataset because gestures in the Text-Gesture generation model are intended to be universal. To remove the information that detailed the intricacies of the speaker's actions, the procedure depicted in Figure 3.1 was taken.

3.4.1 Assessment of Speaker-Specific Gestures

Frames, words, and salient moments from Open Pose were included in the Gestures Data, which concentrated on a single speaker. The key to the suggested research, though, is to create motions that are not particular to any one speaker. Thus, to enable a more detailed examination of the various gesture kinds and the essential points they correspond with, it became necessary to exclude speaker-centric data from the gesture’s dataset. Specifically, the shoulders, arms, wrists, and hands are covered by these important areas.

3.4.2 Extracting Gestures

Words, gesture images, and important points are all included in the Gestures Dataset. Words were taken out in a predetermined amount of time, and every word has a motion attached to it. Given that the speaker is a professor, his dataset contains phrases and gestures that are unique to his position as a teacher. There are probably some words that lecturers use frequently, like greets, introductions, audience interactions, and summaries of earlier lectures. The maximum number of common words were selected, and the corresponding gestures were identified for each word to anonymize the data. Due to a lack of resources, I ran this experiment solely using my dataset.

3.4.3 Selecting the Gestures

Based on research on a single speaker, each word in the Words and Motions dataset has at least three motions associated with it. To optimize the model's training for a specific motion, the Structural Similarity Index (SSIM) was used to identify the optimal gesture image based on similarity value.

3.4.4 Structural Similarity Index

In the field of image processing, the Structural Similarity Index (SSIM) is an essential metric for evaluating image similarity. It is notable as one of the significant milestones in Deep Learning, allowing the investigation of similarity correlations between images. This flexible measure is useful not only for assessing image-focused Deep Learning Models but also for picture production and restoration applications. Using three essential dimensions – Contrast, Structure, and Luminance – SSIM fully captures image similarity. By using a set of weighted functions, deviations in these crucial dimensions are evaluated to calculate the SSIM Index. The resulting SSIM Index values are -1 to 1, where 0 represents total dissimilarity, 1 perfect similarity, and -1 no discernible similarity between two photos. The SSIM Index was utilized in this investigation to assess the degree of similarity [78] between three speaker gesture photographs. Pairwise comparisons between the first and second images, the second and third images, and the third and first images were carried out. The gesture image that was closest to a 1 in the SSIM Index value was chosen. Following that, the model was trained using the relevant key points and terms.

3.4.5 Train and Test Split

The extensively used Train-Test Split methodology is used to evaluate the performance of Deep Learning and Machine Learning Algorithms. This simple process consists of dividing the entire dataset into two halves. The Deep Learning Model is trained using the first,

sometimes referred to as the training set. The results of the second, known as the test set, are used to assess the model's performance because it is not viewed by the model during training. A train-test split ratio of 80:20 is used in this study, meaning that 80% of the data is used to train the model and the remaining 20% for testing.

3.5 Proposed Text-to-Gesture Model

The suggested model makes use of two well-known Deep Learning algorithms for gesture creation, which are well-known for their effectiveness in a range of real-world situations. In particular, the Long Short Term Memory network (LSTM) and the Convolutional Neural Network (CNN) are used. Text data is processed by CNN as a series of inputs, which then turn into word vectors. To record and transmit time series data, five up and five down block operations are carried out [79]. A three-layered LSTM network with a dense layer and a flattened layer comes next. Gestures are ultimately produced using 60 key-point joint coordinates shown in Figure 3.2.

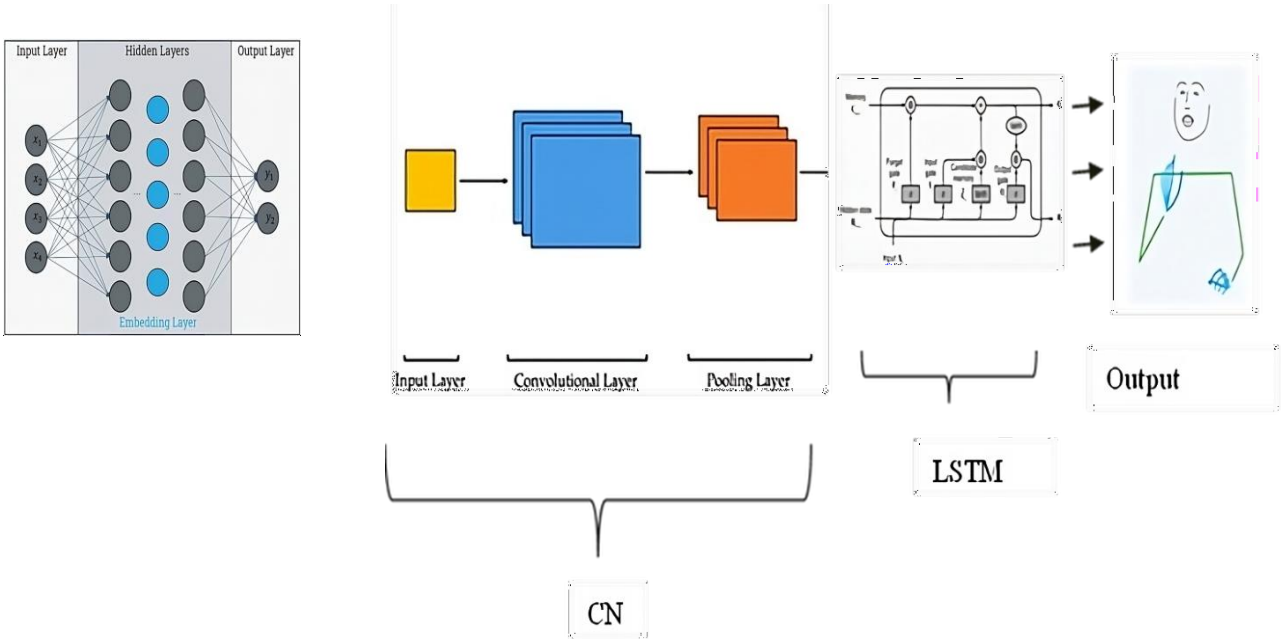

Figure 3.2: Proposed Model

3.5.1 Word Embedding

The translation of textual terms into vectors is an essential step in the domain of Deep Learning and Machine Learning Models when working with text and semantic data. Machine learning models can successfully understand meaning and context when words are represented in a continuous vector space, which offers a mapping from higher dimensions to lower dimensions. The Word2Vec method is used by the suggested Text-to-Gesture model to translate textual words into vectors. To turn a single input word into real-word vectors, this requires using one million English word vectors trained with Wikipedia subword 2017 via the Fast Text Python Library. The goal of Word2Vec, a well-known Deep Learning word embedding model, is to build continuous vector representations based on word occurrences in text. Word2Vec is incorporated into the Text-to-Gesture model that is being presented because of its noteworthy contributions to Natural Language Processing domains like Sentimental Analysis, Text Classification, and Machine Translation.

3.5.2 Convolutional Neural Network

One of the most popular algorithms in many Deep Learning and Machine Learning Models is the Convolutional Neural Network (CNN). This is especially true for applications or issues involving data visualization, image classification, and object or edge recognition in images.

Like an animal's Virtual Cortex, which is made up of several layers of neurons, a CNN also has layers like the Pooling Layer and Convolutional Layer. Either completely or partially connected networks exist between these tiers. CNNs have transformed computer vision by exhibiting remarkable capabilities across a range of tasks. The following are the reasons why the beneficial applications of CNN have been selected for this proposed study.

3.5.3 Operations for Down Sampling and Up-Sampling

Down sampling is a technique used in convolutional neural networks that aims to lessen the feature map's structural complexity. This process is usually performed after the convolutional layer and reduces the number of layers that follow without sacrificing important characteristics. In contrast, down sampling is represented by up sampling, which is often referred to as devolution or Transposed Convolution. Down sampling procedures are essential for collecting layered representations and spatial resolutions in visual data when using convolutional neural networks. Using a convolutional neural network architecture with five down sampling block operations, the feature vector in the proposed research is reduced to 300 × N/32, where N is the total number of frames in the input sequence. The time series data is then reintegrated using five up-sampling block operations, each of which uses a skip connection to send information to send contextual information to the portion of the decoder.

3.5.4 Long Short-Term Memory Network (LSTM)

The Text-to-Gesture Model is based on the three-layered Long Short-Term Memory Network (LSTM) architecture. LSTM is a subtype of Recurrent Neural Network (RNN) that is distinguished by its exceptional accuracy in processing Sequential Data, including audio, video, and semantic text. This improved version finds extensive application in tasks such as text and speech recognition [80], robotics, and handwritten character and digit identification. It is categorically designed to capture and keep long-term dependencies.

The input, forget, and output gates are the three types of gates found in each cell that make up the LSTM Network design:

Input Gate: This component determines which portion of the input should be kept, ignored, or retrained. It oversees controlling the flow of information into the cell state. It is made up of a

Sigmoid Activation Function that multiplies candidate values obtained from the input at hand by element-wise means using weights that have been acquired. The LSTM Network retains relevant details by filtering out irrelevant information using an additional operation. Long-term dependency is retrained with the help of this method.

Forget Gate: Essential to the operation of the Long Short-Term Memory (LSTM), the Forget Gate eliminates extraneous data from the cell and dynamically modifies the memory representation to concentrate on the important components of the input sequence. This skill is essential for handling data that is consecutive.

Output Gate: Choosing which parts of the cell to use to generate the output, the Output Gate oversees transferring crucial data from the present state to the following layer. This gate, which includes a Sigmoid Activation function, conducts element-wise product operations, allowing the LSTM Network to use information selectively to produce the desired output.

3.5.5 Sigmoid and Tanh Activation Functions

To add non-linearity and manage information flow in the LSTM, activation functions are necessary. Tanh and Sigmoid are two of these functions used by the LSTM network. Sigmoid is deliberately used in the input, output, and forget gates. It maps values between 0 and 1, as specified by the equation σ(x) = 1 / (1 + e^−x), acting as a real gate. The LSTM network's hidden and cell states both exhibit non-linearity, which is introduced by the hyperbolic tangent function, Tanh. This function accepts both positive and negative numbers and translates input values to the real range between -1 and 1. Tanh (x) = (e^x − e^−x) / (e^x + e^−x) provides its definition. Tanh and Sigmoid activation functions are essential for preserving specific information, identifying intricate patterns, and adding non-linearity to the LSTM output when processing sequential data. Moreover, these roles support the maintenance of both Long and Short-Term Memory, where Long-Term Memory spans the whole LSTM network while Short-Term Memory is contained in a single cell.

3.5.6 Dense Layer

The 3-layer LSTM network model in the proposed research incorporates a Dense Layer into its Text-to-Gesture Model. Every neuron in this dense layer connects to every other neuron in the layer above. Non-linearity is added to the final output layer by multiplying weights and inputs from the previous layer and then applying activation functions. This layer adjusts bias and weight values to learn complex patterns within sequential data, which makes it an excellent fit. The quantity of neurons in the output layer and this layer match.

3.5.7 Flatten Layer

In a neural network model, the flattened layer serves as a transitional layer that converts multi-dimensional data into a single-dimensional or flat vector. The flattened layer transforms the neural network output into a single 1-D vector by adjusting its shape using variables like width, depth, and height. This layer reshapes and resizes the neural network output, but it doesn't change the output values or carry out intricate operations. The flattened layer plays a crucial role in the suggested text-to-gesture paradigm by translating the output into 60 2-D gestures and producing a series of gestures according to the number of input sequence words.

3.5.8 Process of Generating Gestures

The training of the Text-to-Gesture model uses 90 2-D key-point joint coordinates about the arms, hands, wrists, and shoulders. The mapping that the model does is as follows: G: R $300\times N \rightarrow$ R $90\times N$, where N is the number of words in the input sequence, 300 is the dimensions of the input vector, and 90 is the number of gestures used. Upon obtaining a text sequence as input, the algorithm accurately predicts key points to generate a corresponding gesture image for every word.

3.6 Adam Optimizer

The fine-tuning of model parameters and hyperparameters becomes crucial in Deep Learning and Machine Learning applications that deal with large volumes of high-dimensional data. To achieve this fine-tuning, optimizers are essential. They work to maximize both accuracy and loss by figuring out the best values for model parameters and hyperparameters. The Text-to-Gesture model that is put forward here utilizes adaptive moment estimation or the Adam Optimizer. The AdaGrad and RMSProp algorithms, which are specifically made to handle huge datasets and choose the best learning rates, are combined into one optimizer. The Adam optimizer starts by setting starting weights and then applies the backpropagation technique to compute gradients concerning the loss function. It then calculates the moving average of gradients and bias-correlated averages, repeatedly adjusting weights until the intended result is obtained.

3.7 Summary

An extensive summary of the methods and procedures used to create the Text-to-Gesture model is given in this chapter. It starts by going over the process for eliminating speaker data from the generated dataset. It also explores the strategies used to build the model, tackling the difficulties brought on by the sizable sequential dataset. A brief explanation is given on the application of optimization methods and activation functions. Lastly, for a thorough comprehension, the Text-to-Gesture Generation cycle is described and illustrated through diagrams.

CHAPTER 4

RESULTS AND ANALYSIS

4.1 Overview

This chapter explores the outcomes of the Text-to-Gesture model's implementation, concentrating on the primary job of producing gestures from Urdu text. The paradigm was introduced in Chapter 3. A thorough examination of the effect on gesture quality is provided, along with details on the experimental configurations, platform of implementation, and characteristics that are essential to assessing the model's efficacy. The chapter is structured so that Section 4.3 goes into the analysis and discussion and Section 4.2 gives specifics on the assessment parameters and the outcomes collected. Especially, Section 4.4 presents noteworthy accomplishments and makes analogies with other suggested gesture models. In Section 4.5, the chapter is succinctly summarized.

4.2 Evaluation Parameters

It is impossible to overestimate the importance of evaluation parameters in deep learning and machine learning models since they provide a qualitative way to assess a trained model's performance. These factors allow for comparative research with other state-of-the-art proposed models and provide a way to evaluate the model's efficacy. The evaluation of the Text-to-Gesture model presented here is done according to several parameters, which are explained in the sections that follow.

4.2.1 Percentage of Corrected Key Points (PCK)

One important metric for evaluating the performance of AI models, especially those designed for computer vision applications, is the Percentage of Corrected Key Points (PCK). These models use key points that are taken from photos or moving objects to estimate the stance of people or other moving objects in real-time scenarios.

Within the framework of the suggested Text to Gesture model, PCK was subjected to a series of steps in the evaluation process.

Definition of Threshold Interval: To build a pose estimation, a threshold distance is defined, which gives a certain value for key points that are deemed valid when they are close to the Ground Truth.

Accurate Key-Point Computation: The model computes the difference between the actual Ground Truth and each anticipated Key-Point. The estimated Key-Point is considered accurate if the distance is within the specified range.

Determining the Percentage of Accurately Estimated Crucial Points: The proportion of corrected Key-Points out of all extracted Key-Points is calculated to determine the correctness of the model. The efficacy of the model in relation to the Ground Truth is shown by this percentage, which stands for the PCK.

Establishing an appropriate threshold is a crucial step in computing a PCK value since it defines the region that Key-Points are deemed to have been correctly predicted. A moderate threshold was established during the assessment of the suggested model, and several thresholds were tried

to confirm performance. The big benefit is that only those Key-Points that fit inside a specific range and closely match the Ground Truth are considered as true. Precision is improved by using a close, narrow threshold [81]. α = 0.1 and 0.2 were the threshold values used to assess the suggested model.

Correct Key-Points = No. of Key Points In between Threshold ..... (i)

$$PCK = \frac{No\ .of\ Corrected\ KeyPoints}{Total\ no.of\ KeyPoints} * 100 \quad \text{.....(ii)}$$

4.2.2 Mean Absolute Error (MAE)

One important measure used to verify the performance of the suggested model is the Mean Absolute Error (MAE), which is often accepted and applied for assessing the performance of Deep Learning and Machine Learning models, especially when it comes to regression assignments. In situations when a model is involved in continuous value prediction, MAE is beneficial. The absolute discrepancies between the predicted values and the ground truth, or actual dataset, are computed using this metric. Among its many advantages are its insensitivity to outliers and its ability to provide information on the model's development by averaging the separately computed absolute errors.

The variables 'actual' and 'predicted' indicate the original and predicted values of any marked variable or instance, while 'n' indicates the total number of instances in the dataset. The calculated error value is represented by MAE in this equation. The following procedures were necessary to calculate the MAE for the suggested gesture model:

- Retrieving the real values from every dataset object.
- Figuring out the absolute difference between each data point's actual value and its expected value as predicted by the model.

- Calculating the average by adding up all of the generated values.

It is important to remember that a lower MAE is seen as better when evaluating the MAE value for model evaluation. A reduced MAE indicates that the model's projected values are reasonably close to the ground truth values. Furthermore, it shows that in real-time circumstances, the unit of measurement and the target variable are constant.

$$MAE = (\frac{1}{n}) * \sum |y_{actual} - y_{predicted}| \quad \text{.....(iii)}$$

4.3 Experimental Settings

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

Table 4.1: Experimental Setting

Setting	Value	Description
Batch Size	30	A batch size of 30 is utilized, implying that the model's parameters are updated after processing every 30 samples.
Device	Cuda	The proposed model is trained on a system equipped with a Graphics Processing Unit (GPU) supporting the Cuda framework, facilitating parallel computing.

Epochs	75	100 epochs are employed for fine-tuning the model, indicating that the entire dataset is iteratively passed 75 times through the model during training. This prolonged training period is necessary due to the substantial dataset discussed in Chapter 3.
Models saveInterval	10	The interval for saving the model is set to 10, signifying that weights and bias values are stored on disk after every 10th iteration during training.

4.4 Results and Discussions

Two performance evaluation matrices (PCK)—the mean absolute error (MAE) and the percentage of corrected key points—were used to assess the suggested Text Gesture Model. In the PyTorch [82] environment, the model was implemented. The maximum percentage we were able to obtain for crucial points that were accurately predicted throughout numerous epochs is shown in Figure 4.2. Important information about hands, arms, wrists, and shoulders made up our dataset. Since every joint is next to every other joint, establishing a strict threshold might not lead to any confusion when aiming for optimal performance. The graph exhibits the model's efficiency by plotting the value of PCK versus the number of training epochs.

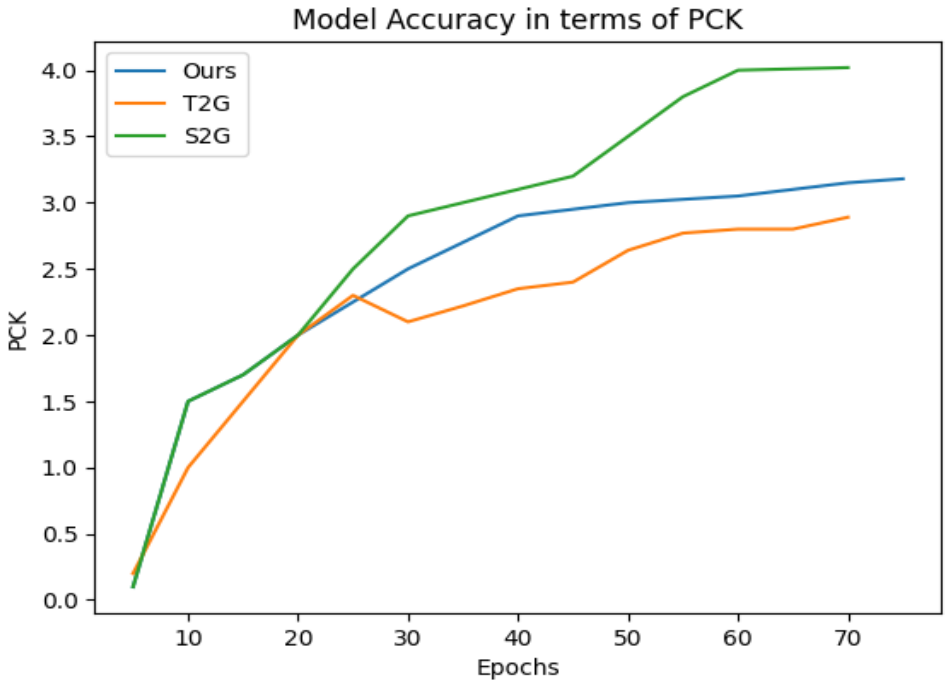

Figure 4.1: PCK of Generated Gestures

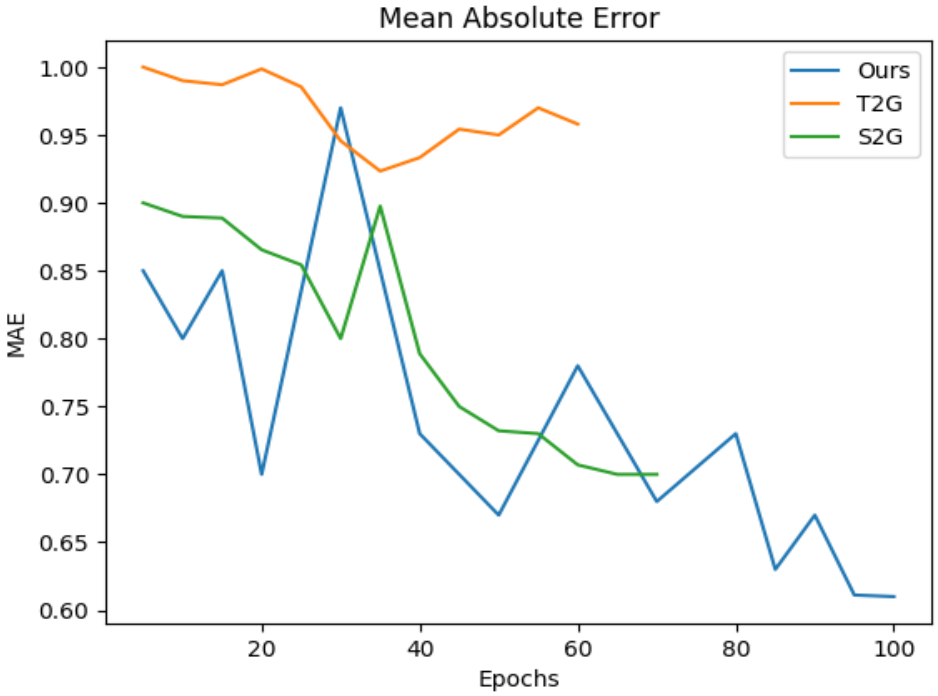

Figure 4.2: MAE of Generated Gestures

The results presented in Figures 4.1 and 4.2 clearly indicate the effectiveness of our proposed Text to Gesture Model in the Urdu language. Achieving an accuracy of 75% based on the percentage of corrected key points is a significant milestone, demonstrating that the model can reliably interpret and generate gestures from Urdu text input. The minimum error rate of 0.3 further supports the precision of our system, indicating that the gestures produced closely match the expected key points extracted from the original videos. This level of accuracy is notable, especially given the nuanced and expressive nature of gestures, which are crucial for conveying meaning and context in communication. The implementation focused on extracting key points from critical parts of the body involved in gesturing, namely the hands, arms, shoulders, elbows, and wrists. This comprehensive approach ensures that the generated gestures are detailed and accurate, capturing the essential movements and positions required for effective communication. The quality of these generated gestures, as evidenced by the high percentage of corrected key points, suggests that the model can produce fluid and natural movements that are representative of human gestural communication. Moreover, the consistency of the model's performance across different samples within the dataset underlines its robustness. The high accuracy achieved highlights the model's capability to generalize well across various inputs, ensuring that the gestures remain coherent and contextually appropriate regardless of the specific text provided. This consistent performance is crucial for applications where reliable and accurate gesture generation is necessary, such as in virtual assistants, language learning tools, and interactive multimedia systems. The results, therefore, affirm the success of our approach in creating a functional and accurate Text to Gesture Model for the Urdu language, paving the way for further advancements in this field. Furthermore, the successful implementation of our Text to Gesture Model in the Urdu language signifies a significant step forward in bridging the gap between textual inputs and non-verbal communication modalities. By accurately translating text into expressive gestures, our model opens new avenues for enhancing human-computer interaction, particularly in multicultural and multilingual contexts where language barriers exist.

Table 4.2: Comparison of Proposed Hybrid DL to Gesture Model

	PCK	**MAE**	**Threshold**	**Language**	**Speaker Specific**
Our Hybrid Model	0.75	0.323	α = 0.1, 0.2	Urdu	No
T2G	0.288	0.958	α = 0.1, 0.2	English	Yes
S2G	0.4	0.707	α = 0.1, 0.2	English	Yes

Table 4.2 illustrates that by using the same parameters quality gestures can be achieved if a model is trained in different languages. Except for our model, every other model made use of the identical benchmark dataset, which included ten speakers' precise motions of every individual speaker the model was trained on every individual speaker also the recorded videos in all datasets were in English language but our dataset was comprised of only one speaker who was an Urdu Language Professor delivering a lecture and the speaker information was kept hidden from the model so our proposed model is not specific to any particular speaker.

4.5 Summary

In Chapter 4, "Results and Analysis," the implementation outcomes of the Text-to-Gesture model are thoroughly examined with an emphasis on the model's primary purpose of producing gestures from Urdu text. To evaluate the model's performance, the chapter offers a

thorough analysis of gesture quality, experimental settings, and important assessment criteria. The painstaking procedures that go into PCK evaluation highlight how useful it is for figuring out whether critical point predictions are accurate. Concurrently, the addition of MAE as a regression task metric offers information about the accuracy of the model. The final section of the chapter focuses on the experimental settings, clarifying important variables like batch size, device specs, epochs, and model save intervals. This ensures that the robustness and efficacy of the model are thoroughly investigated during the phases of experimentation and implementation.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Overview

This study's focus was on investigating the issue of speaker-specific movements and gesture accuracy. The results included the development of a deep learning model and methods that can produce gestures independent of speakers and increase accuracy in terms of PCK value. To the best of our knowledge, the suggested model is unique in that it produces hand gestures against a text input modality and without the assistance of a human. The goal of the suggested Text to Gesture model is to improve and amplify the earlier motions using fresh approaches and methods. A more detailed synopsis of the research is given in Section 5 of this chapter 2. This research's major contributions are outlined in Section 5. Section 3 outlines specific concerns and potential avenues for improving the research. The main goal of this thesis is to create a Text to Gesture generating model with a PCK value that can increase accuracy. Additionally, the impact of the suggested model is examined by a comparative analysis using the most advanced approaches currently accessible.

5.2 Summary of the 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 methods that were currently available trained models on speaker-specific data and generated gestures. This raises the likelihood of a model limitation. The suggested model uses approaches to eliminate speaker information from the dataset to minimize this problem and optimize the model's flexibility. The method utilized in [33] to extract speaker information from the ten people's gestures is shown in Figure 3.4

ii. **High Accuracy:** One key result of deep learning and machine learning is accuracy, which indicates the effectiveness and performance of the model. The suggested gesture generating model consists of two main Deep Learning Algorithms and is based on the Hybrid Deep Learning methodology. Long Short-Term Memory Network with Convolutional Neural Network Its comparison with other cutting-edge methods of manufacturing movements made evident how accurate the model described in this research thesis was. The Mean Absolute Error (MAE) has been minimized and the PCK value has increased by up to 10% with the proposed model. Sections of the preceding chapter demonstrate the effects of employing a sequential Deep Learning Algorithm (LSTM) on the obtained results.

iii. **Gestures against Urdu Text:** Text is extremely important in many domains, such as documentation and knowledge sharing. Interaction between natural and artificial entities, such as humans and computers, depends heavily on hand gestures. For many practical applications, it is consequently necessary to generate hand movements from text. An immersive method of enabling user-computer interaction is through gestures against words. Existing models and methods for generating gestures rely on vocal input; nevertheless, these schemes typically don't offer many benefits. As a result, the model used in this study is based on text input and generates high-quality motions.

Since deep learning and artificial intelligence approaches have the potential to reduce the gap between humans and machines, the suggested text-to-gesture model is based on a hybrid deep learning approach and has outperformed other methods in terms of results. Undoubtedly, a great deal of work has gone into this research, but there are still a number of different approaches to investigate it

5.3 Applications

There are several uses for artificial intelligence in every aspect of daily life. The suggested study is within the purview of artificial intelligence, enabling its use in a variety of real-world contexts. Following is a list of a few of them:

i. **Translating Sign Language:** Hard-of-hearing groups can benefit from support using a Text to Gesture, and deaf individuals can be accommodated in an inclusive atmosphere. This model offers implementation for the conversion [83] of text to signs.

ii. **Visual characters and Assistants:** In virtual reality (VR) and [84] augmented reality (AR) environments, gesture models are widely used to teach animated characters and avatars to match textual content. This can enable the greatest amount of human-virtual assistant engagement.

iii. **Robotics:** The gesture model can be implemented and integrated into artificial robots to enable efficient human-computer interaction by interpreting textual commands and matching the right gesture to each word.

iv. **Modern Assistive Technological Systems:** People manage with assistive devices [85] that serve as their personal assistants. These devices can be integrated with a gesture generation model, which translates text messages into gestures that allow users to interact with their surroundings naturally.

v. **Healthcare:** By precisely executing actions, gesture models can be included and implanted into rehabilitation activities to offer instantaneous instruction.

vi. **Entertainment:** Text to Gesture models can be used in games and other entertaining applications that convey stories [86] and support young learners in developing in a productive and engaging setting.

vii. **Artificial Intelligence-based Interfaces:** Sentiment analysis [87] is applied to the context and exchanges between the user and an artificially intelligent tool, such as chat-gpt, to enable expressive use of AI.

viii. **Story Telling:** When presenting stories, artificial gestures can be employed, with text input provided via storybooks.

5.4 Limitation

Although the methodology for gesture generation provided in this research thesis produces high-quality gestures, it has several drawbacks. Key points on the wrists, arms, hands, and shoulders were among the limited gesture characteristics in the dataset utilized to train the model. The input text contained only Urdu language terms. These restrictions may encourage more research directions to be investigated.

5.5 Future Work

There are no restrictions on the field of study. The only subject of computer science that offers researchers and academics a variety of avenues for investigation is artificial intelligence. The research described in this thesis can also be expanded in several ways to enhance the model's performance and gesture quality. Below is a discussion of a few of them:

i. **Language:** The suggested approach can generate gestures in opposition to the Urdu language-based text input modality. By providing the model with words from any language other than Urdu as input, this work can be improved. This can be accomplished by using the dataset of characters in that language to train the model. It should be noted that gestures from various languages can differ from one another.

ii. **Including Facial Expressions with Hand Gestures:** Given that it contains Key-Points about arms, wrists, hands, and shoulders, the Proposed Gesture Model is trained in a limited set of movements. This gives guidance on how to enhance the gesture model and add facial expressions to the dataset. For instance, if a positive word is used, its context can be changed, and NLP tasks can be carried out to generate a facial smile that corresponds to the movement of that text. In this regard, a few researchers have suggested certain methods.

iii. **Evaluation Measures:** Since percentage of corrected key points (PCK) is the most widely used performance evaluation metric for joint key points data, the proposed Text to Gesture Model's performance is assessed using PCK. More performance metrics that assess a gesture model's effectiveness can be created and improved to advance study in this area. Multi-model performance evaluation will rise as a result.

iv. **Real-Time Gestures:** The gesture model presented in this study generates gestures using a supplementary dataset after the model has been trained. This provides a novel avenue for research, and models can be created and enhanced to generate gestures in response to real-time input. Since this behavior is now being examined, numerous initiatives have been made in this regard.

REFERENCES

[1] Kucherenko, T., Hasegawa, D., Kaneko, N., Henter, G. E., & Kjellström, H. (2021). Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. *International Journal of Human–Computer Interaction*, *37*(14), 1300-1316.

[2] Ferstl, Y., Neff, M., & McDonnell, R. (2021). ExpressGesture: Expressive gesture generation from speech through database matching. *Computer Animation and Virtual Worlds*, *32*(3-4), e2016.

[3] Liu, Y., Mohammadi, G., Song, Y., & Johal, W. (2021, November). Speech-based gesture generation for robots and embodied agents: A scoping review. In *Proceedings of the 9th International Conference on Human-Agent Interaction* (pp. 31-38).

[4] Kucherenko, T., Jonell, P., Van Waveren, S., Henter, G. E., Alexandersson, S., Leite, I., & Kjellström, H. (2020, October). Gesticulator: A framework for semantically aware speech-driven gesture generation. In *Proceedings of the 2020 international conference on multimodal interaction* (pp. 242-250).

[5] Kucherenko, T., Nagy, R., Yoon, Y., Woo, J., Nikolov, T., Tsakov, M., & Henter, G. E. (2023, October). The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the 25th International Conference on Multimodal Interaction* (pp. 792-801).

[6] Ferstl, Y., Neff, M., & McDonnell, R. (2020). Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics*, *89*, 117-130.

[7] Maghoumi, M., Taranta, E. M., & LaViola, J. (2021, April). DeepNAG: Deep non-adversarial gesture generation. In *26th International Conference on Intelligent User Interfaces* (pp. 213-223).

[8] Ishi, C. T., Machiyashiki, D., Mikata, R., & Ishiguro, H. (2018). A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters*, *3*(4), 3757-3764.

[9] Huenerfauth, M. (2008). Spatial, temporal, and semantic models for American Sign Language generation: implications for gesture generation. *International Journal of Semantic Computing*, *2*(01), 21-45.

[10] Wang, J., Zhang, L., Wang, C., Ma, X., Gao, Q., & Lin, B. (2020). Device-free human gesture recognition with generative adversarial networks. *IEEE Internet of Things Journal*, *7*(8), 7678-7688.

[11] Nagy, R., Kucherenko, T., Moell, B., Pereira, A., Kjellström, H., & Bernardet, U. (2021). A framework for integrating gesture generation models into interactive conversational agents. *arXiv preprint arXiv:2102.12302*.

[12] Kucherenko, T., Wolfert, P., Yoon, Y., Viegas, C., Nikolov, T., Tsakov, M., & Henter, G. E. (2023). Evaluating gesture-generation in a large-scale open challenge: The GENEA Challenge 2022. *arXiv preprint arXiv:2303.08737*.

[13] Cui, R., Cao, Z., Pan, W., Zhang, C., & Wang, J. (2019). Deep gesture video generation with learning on regions of interest. *IEEE Transactions on Multimedia*, *22*(10), 2551-2563.

[14] Zhi, Y., Cun, X., Chen, X., Shen, X., Guo, W., Huang, S., & Gao, S. (2023). Livelyspeaker: Towards semantic-aware co-speech gesture generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 20807-20817).

[15] Taranta, E. M., Maghoumi, M., Pittman, C. R., & LaViola Jr, J. J. (2016, October). A rapid prototyping approach to synthetic data generation for improved 2D gesture recognition. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (pp. 873-885).

[16] Kang, J., Zhong, K., Qin, S., Wang, H., & Wright, D. (2013). Instant 3D design concept generation and visualization by real-time hand gesture recognition. *Computers in Industry*, *64*(7), 785-797.

[17] Wu, B., Liu, C., Ishi, C. T., Shi, J., & Ishiguro, H. (2023). Extrovert or Introvert? GAN-Based Humanoid Upper-Body Gesture Generation for Different Impressions. *International Journal of Social Robotics*, 1-16.

[18] Chu, M., & Kita, S. (2016). Co-thought and co-speech gestures are generated by the same action generation process. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(2), 257.

[19] Teshima, H., Wake, N., Thomas, D., Nakashima, Y., Kawasaki, H., & Ikeuchi, K. (2023). ACT2G: Attention-based Contrastive Learning for Text-to-Gesture Generation. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, *6*(3), 1-17.

[20] Kucherenko, T., Hasegawa, D., Kaneko, N., Henter, G. E., & Kjellström, H. (2021). Moving fast and slow: Analysis of representations and post-processing in speech-

driven automatic gesture generation. *International Journal of Human–Computer Interaction*, *37*(14), 1300-1316.

[21] Kucherenko, T. (2021). *Developing and evaluating co-speech gesture-synthesis models for embodied conversational agents* (Doctoral dissertation, KTH Royal Institute of Technology).

[22] Kucherenko, T., Nagy, R., Neff, M., Kjellström, H., & Henter, G. E. (2021). Multimodal analysis of the predictability of hand-gesture properties. *arXiv preprint arXiv:2108.05762*.

[23] Kucherenko, T., Jonell, P., Yoon, Y., Wolfert, P., & Henter, G. E. (2020). The GENEA Challenge 2020: Benchmarking gesture-generation systems on common data.

[24] Nagy, R., Kucherenko, T., Moell, B., Pereira, A., Kjellström, H., & Bernardet, U. (2021). A framework for integrating gesture generation models into interactive conversational agents. *arXiv preprint arXiv:2102.12302*.

[25] Kucherenko, T., Hasegawa, D., Kaneko, N., Henter, G. E., & Kjellström, H. (2019, May). On the Importance of Representations for Speech-Driven Gesture Generation. In *AAMAS* (Vol. 19, pp. 2072-2074).

[26] Thangthai, A., Thangthai, K., Namsanit, A., Thatphithakkul, S., & Saychum, S. (2020). The Nectec gesture generation system entry to the GENEA Challenge 2020. In *Proc. GENEA Workshop. https://doi. org/10.5281/zenodo* (Vol. 4088629).

[27] Kucherenko, T., Hasegawa, D., & Kaneko, N. (2021). Gustav Eje Henter, and Hedvig Kjellström. 2021. Moving fast and slow: Analysis of representations and postprocessing in speech-driven automatic gesture generation. *International Journal of Human–Computer Interaction*, *37*(14), 1300-1316.

[28] Jonell, P., Moëll, B., Håkansson, K., Henter, G. E., Kucherenko, T., Mikheeva, O., ... & Beskow, J. (2021). Multimodal capture of patient behaviour for improved detection of early dementia: clinical feasibility and preliminary results. *Frontiers in Computer Science*, *3*, 642633.

[29] Jonell, P. (2022). *Scalable Methods for Developing Interlocutor-aware Embodied Conversational Agents: Data Collection, Behavior Modeling, and Evaluation Methods* (Doctoral dissertation, KTH Royal Institute of Technology).

[30] Teshima, H., Wake, N., Thomas, D., Nakashima, Y., Kawasaki, H., & Ikeuchi, K. (2023). ACT2G: Attention-based Contrastive Learning for Text-to-Gesture Generation. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, *6*(3), 1-17.

[31] Fares, M., Pelachaud, C., & Obin, N. (2023). TranSTYLer: Multimodal Behavioral Style Transfer for Facial and Body Gestures Generation. *arXiv preprint arXiv:2308.10843*.

[32] Alexanderson, S., Nagy, R., Beskow, J., & Henter, G. E. (2023). Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, *42*(4), 1-20.

[33] Liu, H., Zhu, Z., Becherini, G., Peng, Y., Su, M., Zhou, Y., ... & Black, M. J. (2023). EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Masked Audio Gesture Modeling. *arXiv preprint arXiv:2401.00374*.

[34] Yang, S., Xu, Z., Xue, H., Cheng, Y., Huang, S., Gong, M., & Wu, Z. (2024). Freetalker: Controllable Speech and Text-Driven Gesture Generation Based on Diffusion Models for Enhanced Speaker Naturalness. *arXiv preprint arXiv:2401.03476*.

[35] Oralbayeva, N., Aly, A., Sandygulova, A., & Belpaeme, T. (2023). Data-Driven Communicative Behaviour Generation: A Survey. *ACM Transactions on Human-Robot Interaction*.

[36] Woo, J., Pelachaud, C. I., & Achard, C. (2021, October). Creating an interactive human/agent loop using multimodal recurrent neural networks. In *WACAI 2021*.

[37] Liu, C. (2023). *Speech-Driven Gesture Generation of Social Robot and Embodied Agents* (Doctoral dissertation, UNSW Sydney).

[38] Tuyen, N. T. V., & Celiktutan, O. (2022, March). Context-aware human behaviour forecasting in dyadic interactions. In *Understanding Social Behavior in Dyadic and Small Group Interactions* (pp. 88-106). PMLR.

[39] Valle-Pérez, G., Henter, G. E., Beskow, J., Holzapfel, A., Oudeyer, P. Y., & Alexanderson, S. (2021). Transflower: probabilistic autoregressive dance generation with multimodal attention. *ACM Transactions on Graphics (TOG)*, *40*(6), 1-14.

[40] Spitale, M., & Matarić, M. J. (2021). Toward Automated Generation of Affective Gestures from Text: A Theory-Driven Approach. *arXiv preprint arXiv:2103.03079*.

[41] Fares, M., Pelachaud, C., & Obin, N. (2023). META4: semantically-aligned generation of metaphoric gestures using self-supervised text and speech representation. *arXiv preprint arXiv:2311.05481*.

[42] Bogaers, A., Yumak, Z., & Volk, A. (2020, October). Music-driven animation generation of expressive musical gestures. In *Companion Publication of the 2020 International Conference on Multimodal Interaction* (pp. 22-26).

[43] Fares, M., Pelachaud, C., & Obin, N. (2023). ZS-MSTM: Zero-Shot Style Transfer for Gesture Animation driven by Text and Speech using Adversarial Disentanglement of Multimodal Style Encoding. *arXiv preprint arXiv:2305.12887*.

[44] Zhang, F., Wang, Z., Lyu, X., Zhao, S., Li, M., Geng, W., ... & Li, S. (2023). Speech-driven Personalized Gesture Synthetics: Harnessing Automatic Fuzzy Feature Inference. *Authorea Preprints*.

[45] Ng, E., Romero, J., Bagautdinov, T., Bai, S., Darrell, T., Kanazawa, A., & Richard, A. (2024). From Audio to Photoreal Embodiment: Synthesizing Humans in Conversations. *arXiv preprint arXiv:2401.01885*.

[46] Korzun, V., Gadecky, D., Berzin, V., & Ilin, A. (2022). Speaker-agnostic mouth blendshape prediction from speech. *Computational Linguistics and Intellectual Technologies*, *21*, 323332.

[47] Voß, H., & Kopp, S. (2023). AQ-GT: a Temporally Aligned and Quantized GRU-Transformer for Co-Speech Gesture Synthesis. *arXiv preprint arXiv:2305.01241*.

[48] de Wit, J., Vogt, P., & Krahmer, E. (2023). The design and observed effects of robot-performed manual gestures: A systematic review. *ACM Transactions on Human-Robot Interaction*, *12*(1), 1-62.

[49] Windle, J., Taylor, S., Greenwood, D., & Matthews, I. (2022). Arm motion symmetry in conversation. *Speech Communication*, *144*, 75-88.

[50] Sha, T., Zhang, W., Shen, T., Li, Z., & Mei, T. (2023). Deep Person Generation: A Survey from the Perspective of Face, Pose, and Cloth Synthesis. *ACM Computing Surveys*, *55*(12), 1-37.

[51] Jonell, P., Deichler, A., Torre, I., Leite, I., & Beskow, J. (2021). Mechanical Chameleons: Evaluating the effects of a social robot's non-verbal behavior on social influence. *arXiv preprint arXiv:2109.01206*.

[52] Fan, Y., Lin, Z., Saito, J., Wang, W., & Komura, T. (2022). Joint audio-text model for expressive speech-driven 3d facial animation. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, *5*(1), 1-15.

[53] Yoon, Y., Cha, B., Lee, J. H., Jang, M., Lee, J., Kim, J., & Lee, G. (2020). Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, *39*(6), 1-16.

[54] Asakawa, E., Kaneko, N., Hasegawa, D., & Shirakawa, S. (2022). Evaluation of text-to-gesture generation model using convolutional neural network. *Neural Networks*, *151*, 365-375.

[55] Teshima, H., Wake, N., Thomas, D., Nakashima, Y., Kawasaki, H., & Ikeuchi, K. (2023). ACT2G: Attention-based Contrastive Learning for Text-to-Gesture Generation. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, *6*(3), 1-17.

[56] Khati, U., Singh, P., & Shankar, A. (2020, June). Text Generation through Hand Gesture Recognition. In *Proceedings of the International Conference on Innovative Computing & Communications (ICICC)*.

[57] Ali, G., Lee, M., & Hwang, J. I. (2020). Automatic text-to-gesture rule generation for embodied conversational agents. *Computer Animation and Virtual Worlds*, *31*(4-5), e1944.

[58] Ishi, C. T., Machiyashiki, D., Mikata, R., & Ishiguro, H. (2018). A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters*, *3*(4), 3757-3764.

[59] Gao, N., Zhao, Z., Zeng, Z., Zhang, S., Weng, D., & Bao, Y. (2024). GesGPT: Speech Gesture Synthesis With Text Parsing from ChatGPT. *IEEE Robotics and Automation Letters*.

[60] Nyatsanga, S., Kucherenko, T., Ahuja, C., Henter, G. E., & Neff, M. (2023, May). A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. In *Computer Graphics Forum* (Vol. 42, No. 2, pp. 569-596).

[61] Ghorbani, S., Ferstl, Y., Holden, D., Troje, N. F., & Carbonneau, M. A. (2023, February). ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. In *Computer Graphics Forum* (Vol. 42, No. 1, pp. 206-216).

[62] Saund, C., & Marsella, S. (2021). Gesture generation. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition* (pp. 213-258).

[63] Ferstl, Y., Neff, M., & McDonnell, R. (2021). ExpressGesture: Expressive gesture generation from speech through database matching. *Computer Animation and Virtual Worlds*, *32*(3-4), e2016.

[64] Nagy, R., Kucherenko, T., Moell, B., Pereira, A., Kjellström, H., & Bernardet, U. (2021). A framework for integrating gesture generation models into interactive conversational agents. *arXiv preprint arXiv:2102.12302*.

[65] Kim, H. H., Ha, Y. S., Bien, Z., & Park, K. H. (2012). Gesture encoding and reproduction for human-robot interaction in text-to-gesture systems. *Industrial Robot: An International Journal*, *39*(6), 551-563.

[66] Yang, S., Xu, Z., Xue, H., Cheng, Y., Huang, S., Gong, M., & Wu, Z. (2024). Freetalker: Controllable Speech and Text-Driven Gesture Generation Based on Diffusion Models for Enhanced Speaker Naturalness. *arXiv preprint arXiv:2401.03476*.

[67] Wolfert, P., Robinson, N., & Belpaeme, T. (2022). A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems*, *52*(3), 379-389.

[68] Teshima, H., Wake, N., Thomas, D., Nakashima, Y., Baumert, D., Kawasaki, H., & Ikeuchi, K. (2022, January). Integration of gesture generation system using gesture library with DIY robot design kit. In *2022 IEEE/SICE International Symposium on System Integration (SII)* (pp. 361-366). IEEE.

[69] Thangthai, A., Thangthai, K., Namsanit, A., Thatphithakkul, S., & Saychum, S. (2020). The Nectec gesture generation system entry to the GENEA Challenge 2020. In *Proc. GENEA Workshop. https://doi. org/10.5281/zenodo* (Vol. 4088629).

[70] Harz, L., Voß, H., & Kopp, S. (2023, October). FEIN-Z: Autoregressive Behavior Cloning for Speech-Driven Gesture Generation. In *Proceedings of the 25th International Conference on Multimodal Interaction* (pp. 763-771).

[71] Yoon, Y., Park, K., Jang, M., Kim, J., & Lee, G. (2021, October). Sgtoolkit: An interactive gesture authoring toolkit for embodied conversational agents. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (pp. 826-840).

[72] Liu, X., Wu, Q., Zhou, H., Du, Y., Wu, W., Lin, D., & Liu, Z. (2022). Audio-Driven Co-Speech Gesture Video Generation. *Advances in Neural Information Processing Systems*, *35*, 21386-21399.

[73] Fares, M., Pelachaud, C., & Obin, N. (2023). Zero-shot style transfer for gesture animation driven by text and speech using adversarial disentanglement of multimodal style encoding. *Frontiers in Artificial Intelligence*, *6*, 1142997.

[74] Bhattacharya, U., Rewkowski, N., Banerjee, A., Guhan, P., Bera, A., & Manocha, D. (2021, March). Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)* (pp. 1-10). IEEE.

[75] Spitale, M., & Matarić, M. J. (2021). Toward Automated Generation of Affective Gestures from Text: A Theory-Driven Approach. *arXiv preprint arXiv:2103.03079*.

[76] Xu, C., Yan, J., Yang, Y., & Deng, C. (2023). Implicit Compositional Generative Network for Length-Variable Co-Speech Gesture Synthesis. *IEEE Transactions on Multimedia*.

[77] Taranta, E. M., Maghoumi, M., Pittman, C. R., & LaViola Jr, J. J. (2016, October). A rapid prototyping approach to synthetic data generation for improved 2D gesture recognition. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (pp. 873-885).

[78] Chen, J., Liu, Y., Wang, J., Zeng, A., Li, Y., & Chen, Q. (2024). Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. *arXiv preprint arXiv:2401.04747*.

[79] Rebol, M., Güti, C., & Pietroszek, K. (2021, March). Passing a non-verbal turing test: Evaluating gesture animations generated from speech. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)* (pp. 573-581). IEEE.

[80] Wu, B., Liu, C., Ishi, C. T., Shi, J., & Ishiguro, H. (2023). Extrovert or Introvert? GAN-Based Humanoid Upper-Body Gesture Generation for Different Impressions. *International Journal of Social Robotics*, 1-16.

[81] Li, J., Kang, D., Pei, W., Zhe, X., Zhang, Y., Bao, L., & He, Z. (2023). Audio2Gestures: Generating Diverse Gestures From Audio. *IEEE Transactions on Visualization and Computer Graphics*.

[82] Kucherenko, T., Wolfert, P., Yoon, Y., Viegas, C., Nikolov, T., Tsakov, M., & Henter, G. E. (2023). Evaluating gesture-generation in a large-scale open challenge: The GENEA Challenge 2022. *arXiv preprint arXiv:2303.08737*.

[83] Alexanderson, S., Henter, G. E., Kucherenko, T., & Beskow, J. (2020, May). Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum* (Vol. 39, No. 2, pp. 487-496).

[84] Shen, J., Dudley, J., & Kristensson, P. O. (2021, October). Simulating realistic human motion trajectories of mid-air gesture typing. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 393-402). IEEE.

[85] Tripathi, K. M., Kamat, P., Patil, S., Jayaswal, R., Ahirrao, S., & Kotecha, K. (2023). Gesture-to-Text Translation Using SURF for Indian Sign Language. *Applied System Innovation*, *6*(2), 35.

[86] Ji, L., Wei, P., Ren, Y., Liu, J., Zhang, C., & Yin, X. (2023). C2G2: Controllable Co-speech Gesture Generation with Latent Diffusion Model. *arXiv preprint arXiv:2308.15016*.

[87] Ikeuchi, K., Baumert, D., Kudoh, S., & Takizawa, M. (2019). Design of conversational humanoid robot based on hardware independent gesture generation. *arXiv preprint arXiv:1905.08702*.