# A Multi Modal Approach to Speech-to-Sign Language Generation

Thesis submitted in partial fulfillment
of the requirements for the degree of

*Master of Science*
*in*
*Computer Science and Engineering by Research*

by

Mounika Kanakanti
2019900003
mounika.k@research.iiit.ac.in

International Institute of Information Technology, Hyderabad
(Deemed to be University)
Hyderabad - 500 032, INDIA
May 2024

International Institute of Information Technology
Hyderabad, India

# CERTIFICATE

It is certified that the work contained in this thesis, titled " **A Multi Modal Approach to Speech-to-Sign Language Generation** " by **Mounika Kanakanti**, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Dr. Manish Shrivastava

To *my parents* and *friends*

# Acknowledgments

I would first express my heartfelt gratitude to Prof. Manish Shrivastava, my advisor, who played an instrumental role in shaping my academic and personal journey. His faith in my potential was a catalyst for my growth, and he provided me with a chance to pursue my research interests.

My master's journey was transformative, and I am extremely grateful to my advisor for his pivotal role in this transformation.

I sincerely thank my previous colleagues Satish Pendurthi and Gopinadh Lakkoju for believing in me and helping me take my decision to pursue research at IIITH. I want to extend my heartfelt gratitude to Prof. Bapi Raju for his unwavering support, guidance, and his inspiring approach to mentorship, all of which significantly contributed to my academic and personal development. His willingness to invest in my pursuits and provide guidance was instrumental in helping me pursue my research interests, and I am deeply grateful for his generosity.

My sincere thanks to Shantanu, my collaborator and unwavering friend, for his consistent support throughout my master's journey. I am heavily indebted for his steadfast commitment, which has been a cornerstone in overcoming challenges and achieving milestones.

I would like to express my heartfelt appreciation to Sarath for his thought-provoking discussions that helped me gain a clearer perspective on the practical implications of my research. Special thanks go to Sindhu for being an incredibly optimistic and sweet friend. Her positive outlook and constant encouragement has always bolstered my confidence and motivation, especially during challenging times. I would also like to acknowledge Sai for his overarching support and friendship. His encouragement, kindness, and willingness to lend a helping hand have been truly invaluable.

I would also like to thank my labmates Hiranmai and Ashok for being my peers and sharing valuable feedback in all our discussions at IIIT.

To my lockdown buddies, Harshit, Kinal, Shantika, and Ihita, I extend my heartfelt thanks. The memories we created together during the lockdown have brought me joy and helped make campus life even more beautiful.

Finally, I would like to thank all those who have played a part, big or small, in my academic journey. Your contributions and support have been invaluable, and I am deeply grateful for the impact you have had on my life.

# Abstract

Language is a communication system used to share complex thoughts/ideas and is a powerful tool for social cognition. It relies on a multitude of verbal and non-verbal cues to share information. Analyzing the interplay of these language cues within individuals with distinct sensory experiences provides a valuable perspective for comprehending natural languages. This comprehension is achieved by gaining insights into how analogous contextual information is conveyed through varying modalities. Research in these areas is not only of theoretical interest but may also have important practical implications for building more inclusive solutions.

Sign language is a rich form of communication, uniquely conveying meaning through a combination of signs, facial expressions, and body movements. While Natural Language Processing (NLP) has significantly advanced, progress in supporting sign language has been less substantial. To bridge this gap, automatic sign language translation and generation systems offer an efficient and accessible way to facilitate communication between the deaf and hearing communities. Existing research in sign language generation has predominantly focused on text-to-sign pose generation, while speech-to-sign pose generation remains relatively underexplored. Speech-to-sign language generation models can facilitate effective communication between the deaf and hearing communities. In this work, we propose an architecture that utilises prosodic information from speech audio, and semantic context from text to generate sign pose sequences. In our approach, we adopt a multi-tasking strategy that involves an additional task of predicting face expressions in the form of Facial Action Units (FAUs). FAUs capture the intricate facial muscle movements that play a crucial role in conveying specific facial expressions during sign language generation. We train our models on an existing Indian Sign language dataset that contains sign language videos with audio and text translations. To evaluate our models, we report Dynamic Time Warping (DTW) and Probability of Correct Keypoints (PCK) scores. We find that combining prosody and text as input, along with incorporating facial action unit prediction as an additional task, outperforms previous models in both DTW and PCK scores. We also discuss the challenges and limitations of speech-to-sign pose generation models to encourage future research in this domain.

# Contents

# List of Figures

*Chapter 1*

# Introduction

## 1.1 Background & Motivation

Language, as a multifaceted communication system, serves as a powerful tool for the exchange of complex thoughts and ideas, playing a crucial role in social cognition. This intricate system relies on a myriad of verbal and non-verbal cues to convey information, highlighting the interconnectedness of diverse modalities in the communication process. An insightful analysis of the interplay of these language cues becomes particularly valuable when considering individuals with distinct sensory experiences. By delving into how contextual information is conveyed through various modalities, we gain a comprehensive understanding of natural languages.

Sign language, as an exemplar of language diversity, constitutes a rich and unique form of communication. It seamlessly blends together the fluidity of hand movements and gestures, the expressiveness of facial expressions and head movements, and the subtle nuances of body language. It is this harmony of hand movements and expression that makes it complete and effective. Beyond its theoretical significance, research in this realm holds practical implications for developing inclusive solutions that cater to diverse communicative needs. According to the World Health Organization (WHO), over 1.5 billion people, which accounts for approximately 20% of the global population, live with hearing loss, underscoring the importance of accessibility in communication [26]. While the field of Natural Language Processing (NLP) has made remarkable progress in developing language technologies that simplify daily tasks, the advancement in technology to support sign language has not been as substantial [45]. Bridging this gap, automatic sign language translation and generation systems provide an efficient and accessible means of communication between the deaf and the hearing community.

Recent years have seen a surge of interest in sign language technologies, with researchers exploring various computer vision and deep learning approaches to tackle this complex task [28]. While many of these works utilize text or gloss as input for generation tasks, the area of speech-to-sign language generation remains relatively underexplored [28]. Gloss, often used to represent sign language, has been found to lack accuracy in capturing the complete linguistic and expressive aspects of sign language [42, 50]. A study on the Phoenix dataset [5] showed that a significant portion of the data contained

linguistic elements not present in the gloss representation [50]. While text input aids in generating semantic signs, incorporating audio information, especially prosodic elements extracted from speech, provides a more comprehensive input for sign language generation. This inclusion enables a richer output that captures both semantic and expressive aspects, underscoring the importance of including audio data in sign language technologies to enhance naturalness and accuracy [9, 38]. Towards this end, we extract prosody embeddings from the audio using a pre-trained model and pass this to our sign language generation model as input along with text.

Facial expressions stand as another integral component of sign language, capturing subtle emotional nuances and grammatical markers. The intricacies of facial muscle movements necessitate meticulous annotation efforts, posing a bottleneck in the creation of expansive datasets essential for training robust speech-to-sign language generation models. As a result, advancements in automating or semi-automating the annotation process for facial expressions hold promise for mitigating these challenges, paving the way for more extensive and representative datasets. Incorporating audio information and addressing the challenges of annotating facial expressions are crucial for advancing sign language technologies and enhancing the naturalness and accuracy of sign language generation.

In this work, we introduce MultiFacet, an architecture that uses prosodic information derived from speech coupled with semantic information sourced from text. This integrated data serves as the input for generating keypoints pertaining to both facial and hand movements. Furthermore, our approach includes the prediction of Facial Action Units (AUs) within a multi-tasking setup. We evaluate our model using Dynamic Time Warping (DTW) and Probability of Correct Keypoints (PCK) metrics against the existing Indian Sign Language dataset [15] and demonstrate the critical importance of prosody and facial action unit prediction in better sign language generation.

## 1.2 Research Objectives

The research objectives of this study encompass the following:

- To investigate the impact of incorporating prosodic information from speech audio and semantic context from text in the process of generating sign pose sequences.
- To explore the role and significance of predicting Facial Action Units (AUs) as an auxiliary task for generating expressive sign language poses.
- To evaluate the effectiveness of the proposed MultiFacet architecture in enhancing the quality of sign language generation in comparison to existing models.
- To analyze the challenges and limitations encountered while developing speech-to-sign pose generation models and suggest potential directions for future research.

## 1.3 Thesis Outline

The subsequent chapters of this thesis are structured as follows:

1. In Chapter 2, we describe the related works exploring the role of non-manuals in sign language, non-manual recognition in sign language, facial action units, existing sign language generation approaches, co-speech gesture generation and the current evaluation metrics used in sign language generation.

2. Chapter 3 elaborates on our proposed approach for sign language generation.

3. Chapter 4 provides an in-depth analysis of the quantitative and qualitative evaluation of our approach.

4. In Chapter 5, we shed light on the prominent challenges faced by sign language generation models and the limitations inherent in our proposed model.

5. Finally, Chapter 6 brings this thesis to a conclusion.

*Chapter 2*

# Related Works

In this chapter, we begin by introducing the early methods in sign language generation that mainly used text or gloss (a simplified form of sign language) as inputs. However, majority of these approaches have missed capturing the richness of sign language, which involves not only hand movements but also facial expressions and body language. Recognizing this gap, we move on to discuss the essential role of non-manual markers, like facial expressions, in conveying meaning in sign language.

We then continue with a discussion on recognizing these non-manual markers in sign language videos. Challenges in annotating these markers are acknowledged, leading us to consider the inclusion of pre-trained models for annotating the facial expressions, specifically through the recognition of Facial Action Units. Additionally, this chapter incorporates insights from the domain of co-speech gesture generation, where ideas from generating gestures accompanying spoken language can contribute to sign language generation. Furthermore, we recognize the importance of robust evaluation metrics, ensuring a comprehensive assessment of the effectiveness of sign language generation models.

## 2.1 Existing Sign Language Generation Approaches

Majority of the works in sign language generation are based on text or gloss as inputs [28]. [32], a seminal work in this field, generated continuous hand pose sequences using text as input. They proposed two transformer-based architectures T2S(text to sign) and T2G2S(text to gloss to sign) for sign language generation task. Their work helped alleviate the limitations of previous state-of-the-art methods such as [11] that relied on a look-up table to map predicted gloss to isolated 2D skeleton poses. One of their key contributions was that they proposed the use of a counter embedding, normalized using the sequence length, to help identify the sequence end when using the auto-regressive sign-pose decoder in both architectures. In addition, they included future prediction as an augmentation method during training which required the decoder to predict upto 10 Frames from the input of the current timestep. This helped in preventing the model from repeating previous frame's pose to minimize the training mean-squared-error(MSE) loss over keypoint locations in their experiments, producing significantly better results.

While [32] is a great step in the field, it included only a partial representation of sign language, as facial expressions and body language also play a critical role in conveying meaning [13, 27]. In subsequent works, attempts were made to address this limitation by incorporating both manual (hand movements) and non-manual (facial expressions) features into the generation process. However, these endeavors continued to rely on text or gloss as the primary input modality. For instance, [30] employed adversarial training for multichannel sign production, with text as the input source. Similarly, [33] represented sign sequences as skeletal graph structures, utilizing gloss as an intermediary. [41] generated keypoints for hand movements and facial expressions by concatenating embedding outputs from a text encoder and a gloss encoder.

A distinctive approach by [42] involved the generation of Hamburg Notation System (HamNoSys) notation from text, subsequently converted to continuous sign pose sequences. The HamNoSys notation provides a phonetic representation for sign language poses by using components like symmetry operator, non-manual marker, hand shape, hand orientation and hand location which helped generate better outputs for the generation task. While these approaches made strides towards incorporating non-manual features but still lacked the use of prosodic information as input corresponding to the non-manual features in sign language, thereby limiting the richness and naturalness of the generated sign language sequences.

In the realm of generating photo-realistic sign videos, [31, 37] adopted a strategy where they initially generated skeleton poses from text and subsequently generated sign videos conditioned on these poses. [37] used a transformer-based model to translate text to gloss, and then used the predicted gloss labels to retrieve sign-pose skeletons from a motion graph. These sign-poses were then fed to another transformer architecture for generating images sequences. This approach was limited by the available motion graphs, retrieval accuracy and lack of end-to-end optimization for the overall pipeline. [31] overcame this limitation by proposing a new architecture that used a GAN [21] conditioned on the generated sign-pose skeleton sequence and a style image to generate the final image sequence. While effective in certain aspects, this approach, too, relied fundamentally on textual input, thus overlooking the nuanced aspects introduced by prosody and non-manual features in sign language.

It is important to note that the gloss annotations are not always complete or effective representations of sign language. [50] performed an analysis of the PHOENIX-14T dataset [16] which showed that in 23% of the data, the gloss representation did not include any adjectives or adverbs (intensity modifiers) present in the text transcript. For example, the gloss "WOLKE" (CLOUD) represents both "very cloudy" and "slightly cloudy." Recognizing the crucial role of prosody in sign language expression, [50] introduced gloss enhancement strategies, specifically focusing on the incorporation of intensity modifiers in gloss annotations. This innovative approach aimed to address the loss of prosody in gloss representation, thereby contributing to a more nuanced and expressive rendition of sign language. Intensity modifiers, quantifying nouns, adjectives, or adverbs in a sentence (e.g., "very happy" or "little happy"), serve as essential elements in capturing the subtleties of sign language expression.

Recent endeavors, such as those exploring the use of speech melspectrogram inputs to generate hand movements in Indian Sign Language [15], represent a commendable step in the right direction. However, the generation of hand movements alone, although an essential aspect, remains insufficient to capture the full extent of sign language, given the integral role played by facial expressions and other non-manual features. Hence, the need persists for comprehensive approaches that incorporate prosodic information alongside manual and non-manual features to achieve a more authentic and holistic representation of sign language in the generation process.

## 2.2 The Role of Non-Manuals in Sign Language

Non-manual markers in sign languages serve as vital components for conveying grammatical, pragmatic, and discourse information, enhancing the overall clarity and meaning of signed expressions. These markers play a key role in differentiating between various sentence types, such as declarative statements and questions, contributing to the grammatical functions of sign language. For instance, when forming a yes-no question, the use of raised eyebrows and a specific facial expression becomes a distinctive non-manual marker. In contrast, a declarative statement might involve a different facial expression, showcasing the grammatical versatility of these markers.

Moreover, non-manual markers extend their influence to indicate verb agreement, a crucial aspect of conveying grammatical information in sign language sentences. Through specific facial expressions and head movements, signers can convey the agreement between the subject and the verb in a sentence. To illustrate, consider the examples below:

1. ___wh
   SHOP WH
   'Where is the shop?'

2. SHOP POSSESS
   'The shop is here.'

In the interrogative sentence, the signer employs non-manual markers, like raised eyebrows, to form a question seeking information about the shop's location. In contrast, the declarative sentence utilizes a different set of non-manual markers to convey a statement about the shop's existence. This demonstrates how non-manual elements play a crucial role in sign language syntax, helping to structure sentences based on their intended meaning.

### 2.2.1 Pragmatic Functions

Non-manual markers go beyond mere grammatical functions; they also serve essential pragmatic roles in sign language communication. One such function is **emphasis**. Signers can use specific non-manual markers to emphasize particular elements within a sentence, drawing attention to the most im-

portant information. This emphasis contributes to the nuanced expressiveness of sign languages, allowing signers to convey not only the literal meaning of words but also the emotional or contextual significance attached to them.

Additionally, non-manual markers aid in **clarification** by disambiguating signs that may have multiple meanings in different contexts. The ability to clarify meaning through non-manual expressions becomes crucial in situations where signs might be susceptible to interpretation variations. For example, a sign that can represent multiple objects or actions can be disambiguated through accompanying facial expressions or head movements, providing essential context to the viewer.

### 2.2.2 Discourse Markers

In the realm of discourse, non-manual markers play a pivotal role in managing the flow of conversation. **Turn-taking** is facilitated through facial expressions and head movements, signaling when it is one's turn to speak or when a speaker has finished their utterance. This orchestration of conversational dynamics helps maintain a smooth and organized exchange of information, ensuring effective communication within the signing community.

Furthermore, non-manual markers, such as nodding or shaking the head, provide **feedback** to the interlocutor, indicating comprehension or the need for clarification. This non-verbal feedback mechanism enhances the efficiency of communication, allowing signers to gauge the understanding and engagement of their conversation partners. It contributes to the interactive and dynamic nature of sign language conversations, fostering a collaborative and participatory communicative environment.

In essence, the role of non-manual markers extends beyond the structural aspects of sign language sentences. They are indispensable tools for conveying emphasis, disambiguating meaning, and orchestrating discourse dynamics, enriching the communicative experience in sign languages.

## 2.3 Non-Manual Recognition in Sign Language

[39] presented 3D-CNN based multimodal framework for recognition of grammatical errors in continuous signing videos belonging to different sentence types. First, they used 3D-CNN networks to recognise the grammatical elements from manual gestures (hand movements or signs), facial expressions and head movements. Then they employed a sliding window approach to find the correspondences between these modalities to find the grammatical errors. The significance of this approach lies in its ability to holistically capture the simultaneity of sign language, recognizing not only the manual components but also the nuanced non-manual elements that contribute to linguistic expression. This approach enables the model to discern grammatical errors by examining the relationships and synchrony between manual and non-manual components. By doing so, the model can effectively identify instances where the expression or movement does not align grammatically within the signing context.

In a similar vein, [22] shows that the non-manual components (i.e. facial expressions, eyebrow height, mouth, and head orientation) improves the recognition performance in Kazakh-Russian Sign Language.This underscores the broader trend in the literature, where various works, as reviewed by [20], emphasize the crucial role played by non-manual components in both continuous and isolated sign language recognition.

However, a notable challenge in the existing landscape of sign language research is the often limited scale of datasets used in these studies. Many endeavors, despite their valuable contributions, operate with relatively small datasets, which can impact the generalizability of the models developed.

The recent release of a large pretraining dataset for multiple sign languages by [23] is an important step forward in the field of sign language technologies research. The availability of such a dataset allows for the training of deep learning models on a larger and more diverse set of data, leading to improved performance and generalization of sign language models. Deep learning models require large amounts of annotated data and sophisticated techniques to effectively learn the interplay between manual and non-manual elements of sign language and the context in which they are used. Human annotations provide a way to incorporate the linguistic and expressive aspects of sign language into the models, leading to more realistic and effective sign language generation. However, it is important to note that providing further linguistic information, such as annotations of non-manual features, is crucial in order to fully capture the complexity of sign languages. [24] has released linguistic annotations for manual and non-manual components of 2200 ASL continuous signing video corpora. However, providing human annotations for large datasets can be a costly and time-consuming endeavour. For instance, there exists around 37,000 annotations of non-manual elements for approximately 2,200 utterances in the ASLLRP dataset. While the importance of non-manual markers in sign language is undeniable, their annotating them manually poses significant challenges:

- Subjectivity: Interpreting and annotating facial expressions and head movements can be subjective, as the same expression may have different interpretations in various contexts.

- Multimodality: Sign languages are multimodal, combining manual signs, facial expressions, and body movements. Annotating all these components accurately requires expertise and time.

- Limited Resources: Building annotated corpora for sign languages, especially for non-manuals, is resource-intensive and time-consuming.

In order to address this challenge, we utilise facial action units prediction model to obtain weak labels for our dataset and perform a facial action unit prediction as an auxiliary task.Facial action unit (FAU) prediction offers a potential solution to the challenges of annotating non-manual markers in sign language. FAUs are specific facial muscle movements that correspond to different facial expressions. By predicting FAUs, researchers can indirectly capture facial expressions, allowing for more objective and automated annotation of non-manual markers in sign language data.

## 2.4 Co-speech Gesture Generation

Co-speech gesture generation studies have shown the significance of using both speech and text as input for generating semantically relevant and rythmic gestures, In particular, [25] has done an extensive study on co-speech gesture generation comparing rule-based and learning-based methods. In their work, they dive into the comparison between methods that rely on just audio or text vs those that use both. Audio-based generators have the advantage of using the information of prosody and intonation suitable for inferring kinematics, but struggle with absolute pose vital for conveying the meaning properly. Text-based generators, on the other hand, have rich semantic context but fail to capture the rhythm in generation properly since such information isn't always directly available.

Combining both audio and text as input modality has demonstrated great potential in alleviating the limitations listed above and generating significantly better quality gestures. Pioneering work in this field was simultaneously done by 3 works [1, 17, 46] that proposed different architectures for leveraging both modalities in this task. [46] proposed using different encoders for speech, text as well as speaker identity and the resultant embeddings were passed to an auto-regressive decoder for generating sequence of poses. [1] focused on the relationship between latent representations for both speech and accompanying gestures. They showed that the underlying distributions were skewed and proposed using importance sampling to ensure better coverage. In addition, they highlighted that gesture predictions occur at subword level and incorporated this in their approach by modifying the model architecture to perform alignment between encoded sub-words and acoustics using a multi-scale transformer [40]. [17] proposed extracting semantic features from text using BERT [10] and audio features using a convolutional encoder on the Mel-spectograms of the speech input. They used a sliding window approach of the acoustic features to provide past and future context, along with the text embeddings, to the auto-regressive decoder for generating pose sequences. By having the model predict upto three frames consecutively, they were able to enforce better temporal continuity in the predicted gesture sequences. Note all these different works have inspired similar work in the field of sign language generation as well.

In parallel to the research using auto-regressive transformer models, use of motion-graphs for co-speech gesture generation is also being actively explored. A recent work [49] used the StyleGestures [2] model and proposed attributes such as wrist speed, radius and height to generate style signatures for audio signals (embeddings) and gestures respectively. They further extracted rhythm signatures are extracted as bit vectors based on occurrence of words in text and stationary moments in the gesture sequences. Using the audio and text signatures for the input segments, they were able to model gesture generation as optimization task that uses costs based on retrieved motion nodes and distance between adjacent retrieved nodes in the motion graph. Their approach was able to achieve the highest naturalness score in the GENEA 2022 [44] challenge.

## 2.5 Evaluation Metrics for Sign Language Generation

The majority of the works in sign language generation report back translation scores [30, 32, 41]. Back translation involves the process of taking a generated sign language sequence and converting it back into the source language, typically spoken or written language. This back-translated version is then compared to the original source. While back translation scores offer valuable insights into the quality of generated sign language, it's important to note that they carry the risk of error propagation.

Error propagation occurs when initial translation from the source text to sign language using a generative model introduces errors or inaccuracies. The back translation process, similar to the forward pass, is also performed using a generative model and hence not perfect. It corrupts these errors in translation by adding its own when regenerating the source from the generated sequence. This potentially leads to inaccurate and inconsistent analysis of generated sign language and complicates the identification of the root cause of the problem. This makes it challenging to pinpoint where errors originated and also deduce appropriate measures to address them.

Furthermore, sign language, being a rich and expressive form of communication, encompasses non-verbal elements such as facial expressions, body movements, and other cues critical for conveying meaning. These aspects, due to limitations in modeling capacity, may not be fully captured in the back translation process, resulting in a loss of information and the potential for misinterpretations.

To provide a more comprehensive and accurate evaluation of sign language generation models, [15] introduce Dynamic Time Warping (DTW) and Percentage of Correct Keypoints (PCK) scores as additional metrics. These metrics offer insights into the alignment and accuracy of keypoints in generated sign language gestures. They help evaluate the generated output directly against a ground-truth value, thus, overcoming the limitations of back translation. However, these metrics have their own drawbacks which need to be discussed.

The DTW metric computes the optimal alignment of two sequences by taking into account the MSE(mean-squared-error) between keypoint positions in two frames. This can lead to subpar assessment in cases where the keypoints aren't uniformly distributed across features such as faces, hands and body parts. The PCK metric uses a constant radius threshold to define where the predicted keypoint position is sufficiently close to its ground-truth position. This also leads to an incomplete evaluation as it does not factor the range of motion of keypoints belonging to different features. For example, considering keypoints belonging to small features like eyes or fingers, even a small change in positions can bring significant difference to perception as compared to keypoints belonging to shoulder or elbow joints.

Additionally, [25] may provide further insights into the evaluation landscape, emphasizing the importance of using a combination of metrics and human evaluation to thoroughly assess the quality and performance of these models. For subjective evaluation, they propose using two sequences of generations from the same model, one from a relevant input source shown to the participants and the other is randomly chosen. Using generations from the same model helps abate the risk of bias of human-likeness over semantic relevance. In addition, the approach can be used for probing grounding of proposed ap-

proach in different modalities used as input sources for the generation task. For instance, a sign language generation model that takes both audio and image modalities as input can be evaluated on its adherence to each when it comes to human-likeness as well as semantic relevance. For objective evaluation, they re-iterate over the limitations of metrics like PCK and motion properties like acceleration and jerk as good proxies for measuring human-likeness or semantic relevance. Also, they elaborate about how such scores limit the learning capabilities since they expect a one-to-one mapping which is ill-posed for the task of sign language generation. They do explore other studies which propose additional metrics such as Fréchet Inception Distance (FID) and Inception Score but they again only capture appearance and not semantic relevance. They cite [18] to emphasize the poor correlation of the above metrics with human scores in a study done by Geneva on the 2022 challenge [44] submissions.

From the available literature, it becomes clear that there are limitations in picking an appropriate metric for comprehensive and fair evaluation of sign language generation task. For our work, we decide to go ahead with PCK and DTW due to their prevalence in existing literature and direct evaluation using ground-truth instead of model-based metrics which have their own biases and limitations. We report scores using these metrics during our evaluation and follow up with qualitative evaluation to put forth a more comprehensive evaluation.

*Chapter 3*

# MultiFacet Architecture and Approach

Given audio and text inputs, our aim is to generate sequences of sign poses denoted as $\mathbf{S}$, which include both upper body and face keypoints. To accomplish this, we adopt a multi-task learning approach, incorporating a speech encoder, a Facial Action Units decoder, and a sign pose decoder. The overall architecture is illustrated in Figure 3.1.

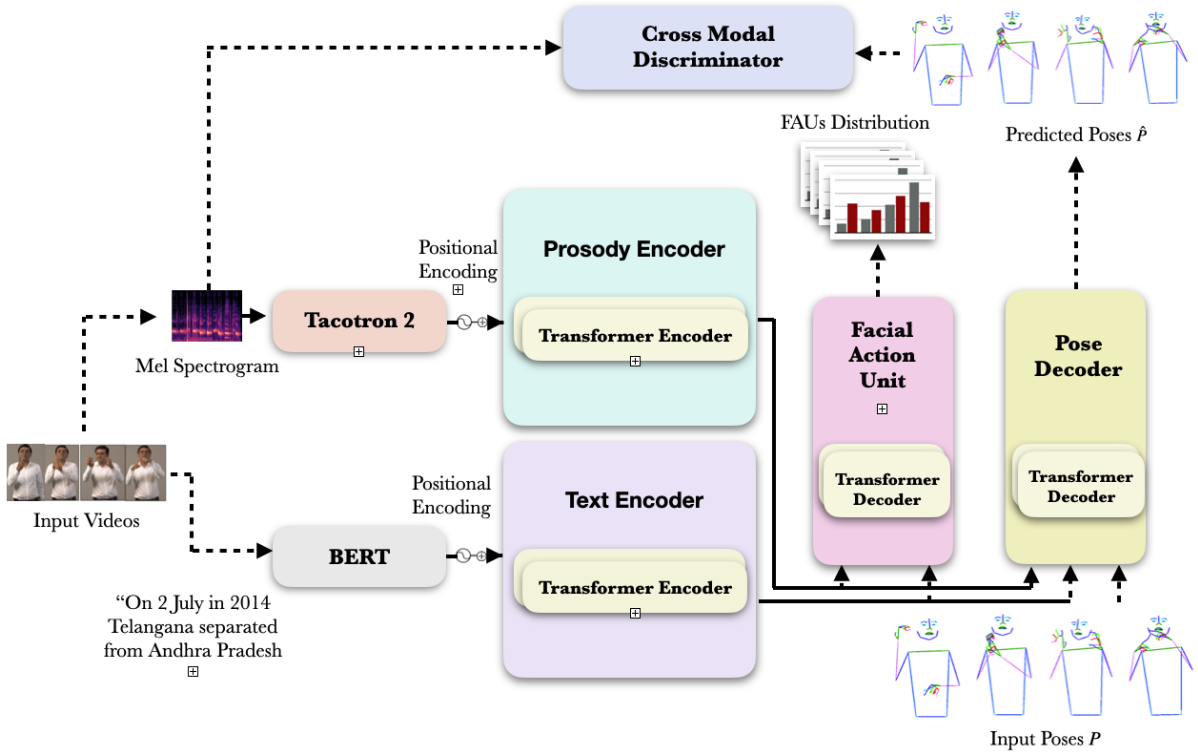


Figure 3.1: The Architecture: We propose a novel architecture to generate sign pose sequences by utilising the prosodic information from speech and semantic context from text. We also incorporate additional components to facilitate rich sign pose generation: (i) Facial Action Unit decoder and (ii) Cross Modal Discriminator.

### 3.0.1 Input Embeddings

We represent the input text as a sequence of tokens $\{x_1, x_2, ..., x_W\}$, and BERT provides the corresponding embeddings $\{e_{x_1}, e_{x_2}, ..., e_{x_W}\}$. To facilitate the generation process, we extract two types of embeddings from the input data: BERT embeddings for text and Tacotron 2 GST [43] encodings for audio. We use the GST model provided by NVIDIA[1] which was pre-trained on *train-clean-100* subset of LibriTTS dataset [48] to represent the expressive features in audio. The main aim of Tacotron 2 GST model behind learning the "style embeddings" was to be able to control synthesis in novel ways, such as varying speed and speaking style – independently of the text content. It is important to ote that the Tacotron 2 GST model includes a reference encoder that takes Mel-spectrograms as input, in addition to the text encoder from Tacotron 2, and outputs a style embedding. This style embedding is then passed to the decoder along with the text for synthesis. In our approach, we pass the Mel-spectrogram to the reference encoder to obtain the style embeddings.

The BERT embeddings, denoted as $\mathbf{E}_{\text{text}}$, capture the semantic information embedded within the text, allowing our model to understand the linguistic context. The shape of the text embeddings is W $\times$ 768.

The Tacotron 2 GST encodings, denoted as $\mathbf{E}_{\text{audio}}$, extract both linguistic content and prosody information from the audio input. The GST model was pretrained on LibriTTS dataset [48] with the objective of learning a large range of acoustic expressiveness. We represent the audio input as a sequence of mel-spectrograms $\{m_1, m_2, ..., m_T\}$, where each mel-spectrogram has T $\times$ 256 dimensions. Tacotron 2 GST [43] provides the corresponding embeddings $\{e_{m_1}, e_{m_2}, ..., e_{m_T}\}$.

### 3.0.2 FAUs Preprocessing

The Facial Action Coding System (FACS) is a comprehensive and standardized method for denoting facial expressions. It is a meticulously designed tool aimed at describing and analyzing nonverbal cues through the precise identification of distinct facial muscle movements. At the core of the FACS system are its Action Units (FAUs), which represent individual facial muscle actions. When combined, these FAUs efficiently portray a wide range of emotions and expressions. The efficacy of FACS has led to its widespread application across various disciplines, including psychology, neuroscience, anthropology, and computer graphics. FACS provides an objective and systematic means to categorize and comprehend facial expressions. A few examples of facial action units include inner brow raiser, upper lid raiser, jaw drop, lip tightener etc.

While FACS offers a robust framework for understanding facial expressions, it is important to note that it generally does not provide information about the degree of muscle activation. In other words, FACS focuses on identifying which facial muscles are involved in an expression but does not quantify the intensity or strength of muscle movements. While there are modifiers that extend this coding system to accommodate intensities as well, we don't consider them in our study due to limited resources and no clear consensus on their use.

[1] `https://github.com/NVIDIA/mellotron/tree/master`

The use of FACS for sign language translation or generation is relatively understudied [7, 8, 35]. One of the primary reasons for its limited use is the costly annotation required for the existing sign language datasets. To overcome this issue, we propose using an existing state-of-the-art model, ME-GraphAU [19], to predict the action units for our chosen dataset and use it as weak-supervision during sign-language generation task. We encourage readers to refer to [19] for details related to architecture, training dataset and output format for the aforementioned model.

The output of the chosen model is noisy and lacks temporal consistency since the prediction occurs on a per-frame basis. Training with such an output would invariably lead to noisy supervision and poor learning on the model's part for the proposed task. As such, we propose a pre-processing pipeline for reducing the noise using the following steps:

- Threshold the output of the model using the probabilities as confidence for each action unit and remove any low confidence predictions.

- For these pruned predictions, we use linear interpolation for estimating their new values.

- Finally, to reduce the remaining noise, we use hanning smoothing over each action unit and get the final output. We use a window length of 11, which corresponds to 0.5 seconds at 24FPS frame-rate of our source videos.

We show an example of the original prediction and output of each step in the above-mentioned pipeline in Figure 3.2.

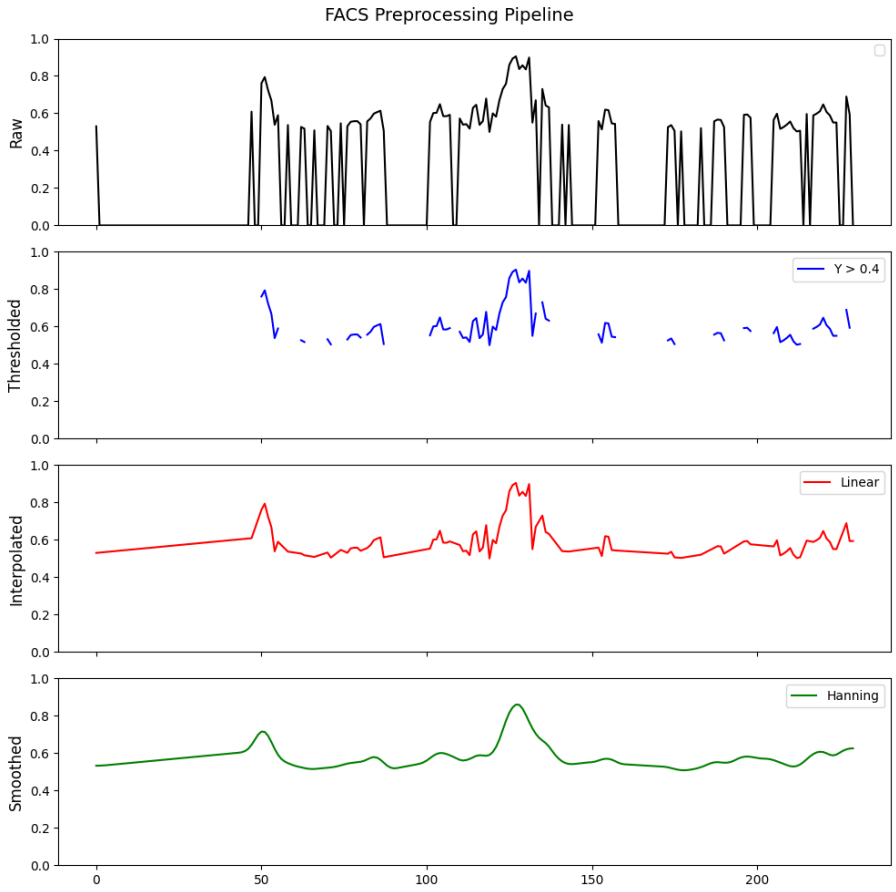


Figure 3.2: Illustration of the Facial Action Units (FAUs) preprocessing pipeline: thresholding using action unit probabilities, linear interpolation, and Hanning smoothing.

Figure 3.3 shows the ground truth facial action units extracted.

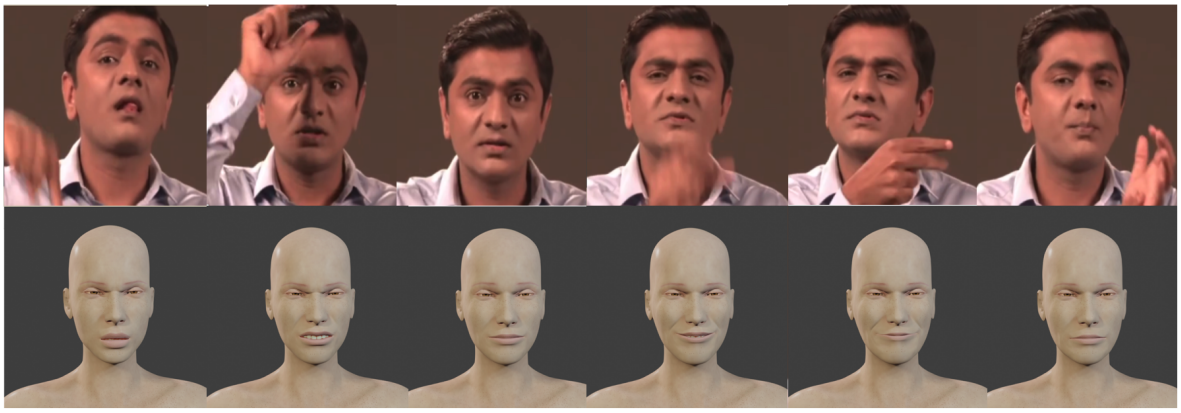

Figure 3.3: Representation of Ground Truth Facial Action Units, generated using Blender [4] for visualization purposes.

### 3.0.3 Model Components

The input embeddings $\mathbf{E}_{\text{text}}$ and $\mathbf{E}_{\text{audio}}$ are then passed to their respective encoders in our model:

1. **Prosody Encoder**: The transformer-based speech encoder, denoted as $E_{\text{speech}}$, processes the Tacotron 2 GST encodings $\mathbf{E}_{\text{audio}}$ to obtain intermediate representations $\mathbf{H}_{\text{speech}}$. This can be expressed as:

$$\mathbf{H}_{\text{speech}} = E_{\text{speech}}(\mathbf{E}_{\text{audio}})$$

2. **FAUs Decoder**: We incorporate the FAUs prediction task as an additional objective to capture facial expressions. The FAUs decoder, denoted as $D_{\text{FAUs}}$, processes the Tacotron 2 GST encodings $\mathbf{E}_{\text{audio}}$ to predict the Facial Action Units, denoted as $\mathbf{FAUs}$. This can be expressed as:

$$\mathbf{FAUs} = D_{\text{FAUs}}(\mathbf{E}_{\text{audio}})$$

Facial AUs is a widely used facial expression coding system that consists of a set of action units that correspond to different facial muscle movements. We use a transformer-based decoder [40] for this task and train it using cross-entropy loss.

$$\mathcal{L}_{\text{FAUs}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{M} y_{n,i}\log(p_{n,i}) \tag{3.1}$$

where $N$ is the number of training examples, $M$ is the number of Facial Action Units, $y_{n,i}$ is the ground-truth label for the $i$-th Facial Action Unit in the $n$-th example (either 0 or 1), and $p_{n,i}$ is the predicted probability for the $i$-th Facial Action Unit in the $n$-th example.

3. **Sign Pose Decoder**: Our sign pose decoder, denoted as $D_{\text{pose}}$, is a transformer-based autoregressive decoder that takes the intermediate representations $\mathbf{H}_{\text{speech}}$ as input to generate the sequence of sign

poses $\mathbf{S}$. The keypoints for each frame in the sign pose sequence are represented as a 3D tensor, with dimensions num_frames $\times$ 85 $\times$ 3. The output of the decoder can be formulated as:

$$\hat{y}_{n,i} = \mathbf{D}_{\text{Pose}}(\mathbf{H}_{\text{speech,n}}, \mathbf{y}_{n,0:i-1}) \tag{3.2}$$

Note that during training, the decoder uses ground-truth poses as input for stability and faster convergence. During inference, the pose inputs to the decoder are its own predictions upto the given timestep.

We use regression loss to train the sign pose decoder, given by:

$$\mathcal{L}_{\text{pose}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{85} \|y_{n,i} - \hat{y}_{n,i}\|^2 \tag{3.3}$$

where $N$ is the number of training examples, $y_{n,i}$ is the ground-truth value of the $i$-th keypoint for the $n$-th example, and $\hat{y}_{n,i}$ is the predicted value of the $i$-th keypoint for the $n$-th example.

4. **Cross-Modal Discriminator** Cross-modal discriminators are a type of deep learning architecture that aims to learn a common representation for different modalities, such as vision and language [21]. These networks leverage the strengths of multiple modalities to improve performance in various tasks, such as cross-modal retrieval and common representation learning.

The motivation behind incorporating the Cross-Modal Discriminator lies in its ability to provide an additional layer of scrutiny. While the regression loss primarily focuses on minimizing the differences between generated and ground-truth sign sequences, the Cross-Modal Discriminator helps to assess how well these sequences align with the characteristics of the input speech. For this purpose, use the same discriminator used by [15] to match the speech segments with corresponding pose sequences.

The primary objective of the Cross-Modal Discriminator is to evaluate the alignment between the provided speech segments and the corresponding generated sign pose sequences. It accomplishes this by comparing the speech representations obtained through the prosody encoder ($H_{\text{speech, n}}$) with the ground-truth ($y_n$) and predicted ($\hat{y}_n$) pose sequences.

The loss for the cross-modal discriminator can be defined as follows:

$$\mathcal{L}_{\text{G}}^{\text{GAN}} = \frac{1}{N}\sum_{n=1}^{N} \log(1 - (\mathbf{D}_{\text{cross-modal}}(\mathbf{H}_{\text{speech, n}}, \hat{\mathbf{y}}_{\text{n}}))) \tag{3.4}$$

$$\begin{aligned}\mathcal{L}_{\text{D}}^{\text{GAN}} = -\frac{1}{N}\sum_{n=1}^{N} &\log((\mathbf{D}_{\text{cross-modal}}(\mathbf{H}_{\text{speech, n}}, \mathbf{y}_{\text{n}}))) \\ &+ \log(1 - (\mathbf{D}_{\text{cross-modal}}(\mathbf{H}_{\text{speech, n}}, \hat{\mathbf{y}}_{\text{n}})))\end{aligned} \tag{3.5}$$

where $\mathbf{D}_{\text{cross-modal}}$ is the cross-modal discriminator. $\mathbf{H}_{\text{speech, n}}$ is the intermediate representation for the $n$-th example obtained by the prosody encoder. Variables $\mathbf{y}_{\text{n}}$ and $\hat{\mathbf{y}}_{\text{n}}$ are the ground-truth and predicted pose sequences respectively. $\mathcal{L}_{\text{D}}^{\text{GAN}}$ and $\mathcal{L}_{\text{G}}^{\text{GAN}}$ are the standard binary cross-entropy loss used for discriminator and generator respectively.

### 3.0.4 Multi-Tasking Setup

Multi-Task Learning has emerged as a powerful paradigm in the realm of deep learning, offering a range of benefits that contribute to enhanced model performance and generalization [6, 29]. In the context of our sign pose generation task, the multi-tasking setup is strategically employed to exploit the synergies among different subtasks, fostering a more robust and versatile learning process. Multi-task learning leverages the inherent relationships among different tasks. In our architecture, the FAUs decoder, speech encoder, and sign pose decoder collectively address diverse aspects of the input data, such as facial expressions, prosody, and sign pose generation. The information gained by each task can be shared and transferred, leading to a more comprehensive understanding of the input. One notable advantage of multi-task learning is its adaptability to challenges associated with weak supervision. This proves particularly valuable in sign language datasets, where obtaining precise annotations for facial expressions might be intricate. By incorporating a task with more readily available annotations (e.g., FAUs prediction), the learning process for another task (e.g., sign pose generation) is guided, showcasing the practicality of multi-task learning in scenarios with varied annotation complexities.

To measure how well our model is doing, we use a weighted sum of losses from individual decoders to compute the overall loss:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{FAUs}} \cdot \mathcal{L}_{\text{FAUs}} + \lambda_{\text{pose}} \cdot \mathcal{L}_{\text{pose}} + \lambda_{\text{discriminator}} \cdot \mathcal{L}_{\text{G}}^{\text{GAN}}$$

where $\lambda_{\text{FAUs}}$, $\lambda_{\text{pose}}$, and $\lambda_{\text{discriminator}}$ are hyperparameters that control the relative importance of the FAUs loss, pose loss, and discriminator loss, respectively.

The weighted sum ensures that the model optimizes its parameters to minimize the combined loss, effectively balancing the objectives of facial expression prediction, sign pose generation, and the alignment of speech segments with corresponding pose sequences. The optimization process involves training the model to minimize the multitasking loss $\mathcal{L}_{\text{total}}$ using gradient-based optimization techniques. In summary, the multi-tasking setup harnesses the benefits of shared learning, regularization, and improved data efficiency to equip our model with a robust capability for sign pose generation. By jointly optimizing multiple interconnected tasks, the model gains a nuanced understanding of sign language inputs, leading to more accurate and context-aware sign pose sequences.

*Chapter 4*

# Performance Evaluation

## 4.1 Experiments

In this chapter, we present a comprehensive overview of the dataset and experiments conducted to evaluate and refine our sign language generation model. We compare our model with two other models, Text2Sign and Speech2Sign, to see how well it performs. To measure our model's performance, we use metrics like Dynamic Time Warping (DTW) and Probability of Correct Keypoints (PCK). These help us understand how well our model aligns signs over time and how accurate its keypoint predictions are. The following chapter then dives into the results and what we have learned from them, giving us a clear picture of where our model excels and where it can be improved. We also conduct ablation analysis, which means breaking down our model into different parts to see how each component contributes to its overall performance. This helps us understand the importance of different modules, like the Facial Action Units (FAUs) decoder. Finally, we share specific details about how our model is built and trained, providing a behind-the-scenes look at the decisions that shape its behavior. Overall, this chapter walks through our experiments, highlighting both challenges and successes in making our sign language model better.

### 4.1.1 Dataset

The dataset used in our study is the continuous Indian Sign Language dataset, which was released by [15]. This dataset contains sign videos along with corresponding audio and text transcription, covering various topics, such as current affairs, sports, and world news. The dataset has a vocabulary of 10k words and comprises of 498 videos with a train-validation-split of 480:9:9. These videos are parsed with a sampling rate of 25 frames per second and their corresponding audio is sampled at 44 KHz. These videos are further split into 9137 segments using timestamps of sentence boundaries in the corresponding subtitles. These different segments have lengths varying from 3 seconds upto 18 seconds with the 90th percentile around 6 seconds. We skip the segments above the maximum length of 6 seconds to avoid unnecessary padding for majority of the sequences and keep the training efficient.

#### 4.1.1.1 Audio Pre-processing

To extract Tacotron [43] GST encodings from the raw audio files, we generate the normalized mel-spectogram as follows:

1. Resample the audio at 24KHz
2. Run preemphasis over the parsed audio waveform to boost high-frequency components.
3. Run Short term Fourier transform (STFT) to generate the spectogram for the waveform with fixed window size.
4. Convert the spectogram to mel-scale using fixed number of mels.
5. Normalize the mel-spectogram using a fixed minimum and max amplitude (in decibels) for scaling and clipping the final values.

After pre-processing, the resultant normalized mel-spectograms are then provided as an input to the Tacotron [43] model and the embeddings from the GST layer are saved for use with our models.

#### 4.1.1.2 Video pre-processing

To represent the sign videos in our analysis, we extracted 3D joint position keypoints using Mediapipe [12]. This process involved detecting $37$ landmark points for the eyes, eyebrows, lips, and face outline, along with $6$ landmark points for the shoulders, elbows, and hips. Additionally, each hand was represented with $21$ landmark points, bringing the total to $85$ keypoints for upper body, hands and face.

The extracted keypoints are noisy and have several issues such as lack of temporal continuity, missed detections due to occlusion or motion blur, as well as incorrect detections particularly for finger joints. Further, in majority of the sign videos, the signer is often not spatially centered in a frame and their extents change as they move around during the course of the video. This leads to a varying offset added to the keypoint locations which isn't related to sign pose sequence and acts as a noise in the learning process. Lastly, the different signers across these videos have different body structures and joint lengths which again isn't relevant to the sign pose sequence and posses a challenge similar to the previous one.

To address these issues, we use the pipeline proposed by [47] which comprises of the following steps:

1. Remove the spatial offset by using the center of the two shoulder keypoints as origin and updating the other keypoints locations to their relative positions.
2. Remove structural differences between signer skeletons by normalizing the bone-lengths using the distance between the two shoulder joints.
3. Use a temporal window to identify outliers for keypoint locations (incorrect detections) and mark them as missing.

4. Use a Gaussian filter for imputing the missing values for keypoint locations.

5. Use 2D-3D skeleton transformation for optimizing the imputed values using backpropagation.

The above steps help circumvent some of the aforementioned issues but there are still several challenges that we discuss in detail in the following chapter.

### 4.1.2 Baseline Models

**Text2Sign** We adopt the progressive transformers introduced by [32] as the foundation of our approach. We use their text-to-sign (T2S) architecture as a baseline for comparison. We extend their proposed architecture to predict 3D keypoints for face and upper body and train them on the aformentioned Indian Sign Language Dataset .

**Speech2Sign** [15] utilised mel spectrograms as input to generate sign pose sequences of hand movements. They incorporate a text decoder and a cross-modal discriminator for learning the correlation between speech and sign pose sequences. We again extend their architecture to generate face and body key points and consider it as another baseline.

### 4.1.3 Evaluation Metrics

**Dynamic Time Warping (DTW)**

Dynamic Time Warping (DTW) [14] is one of the evaluation metrics for speech-to-sign language generation models to assess the alignment between the predicted sign language sequences and the ground truth sign language sequences.

Let $P = (p_1, p_2, \ldots, p_M)$ denote the predicted sign language sequence, where $p_i$ represents the $i$-th pose in the predicted sequence, and $M$ is the length of the predicted sequence. Similarly, let the ground truth sign language sequence be denoted as $G = (g_1, g_2, \ldots, g_N)$, where $g_i$ represents the $i$-th pose in the ground truth sequence, and $N$ is the length of the ground truth sequence.

DTW aims to find an optimal alignment between the sequences $P$ and $G$ by introducing a warping path $W = \{(w_1, w_2, \ldots, w_K)\}$, where $w_k = (i, j)$ denotes the alignment of $p_i$ in the predicted sequence with $g_j$ in the ground truth sequence. The warping path satisfies the conditions: $w_1 = (1, 1)$, $w_K = (M, N)$, and $w_k - w_{k-1} \in \{(1, 0), (0, 1), (1, 1)\}$, allowing for insertions, deletions, and matches between the sequences.

The objective of DTW is to minimize the accumulated cost along the warping path $W$, which is defined by a distance or similarity measure between the individual poses in the sequences. Let $d(p_i, g_j)$ represent the distance between $p_i$ and $g_j$ in the pose space. The accumulated cost $C(W)$ along the warping path $W$ is given by:

$$C(W) = \sum_{k=1}^{K} d(p_{w_k}, g_{w_k})$$

To compute the final DTW score, we aim to find the optimal warping path $W^*$ that minimizes the accumulated cost $C(W)$:

$$DTW(P, G) = \min_{W} C(W)$$

The DTW score provides a measure of the alignment between the predicted and ground truth sign language sequences, considering the temporal differences and variations in the movement patterns. A lower DTW score indicates a better alignment and higher similarity between the sequences.

**Probability of Correct Keypoints (PCK)**

PCK [3, 36] is a widely used evaluation metric to assess the accuracy of pose estimation models. It measures the percentage of correctly predicted keypoints within a certain threshold distance compared to the ground truth keypoints.

Let $G = \{g_1, g_2, ..., g_N\}$ be the set of ground truth keypoints, and $P = \{p_1, p_2, ..., p_N\}$ be the set of predicted keypoints. Each keypoint, $g_i$ or $p_i$, consists of $(x, y, z)$ coordinates representing the position of a particular body part, such as a hand or face.

To compute the PCK score, we need to define a threshold distance $\delta$. For each ground truth keypoint $g_i$, we check if there exists a corresponding predicted keypoint $p_j$ within the threshold distance $\delta$. If such a predicted keypoint exists, and its distance to the ground truth keypoint is less than or equal to $\delta$, we consider it as a correct prediction.

Mathematically, the PCK score can be computed as follows:

$$PCK = \frac{1}{N} \sum_{i} \delta(g_i, p_i)$$

where $N$ is the total number of keypoints, and $\delta(g_i, p_i)$ is an indicator function defined as:

$$\delta(g_i, p_i) = \begin{cases} 1, & \text{if } ||g_i - p_i|| \leq \delta \\ 0, & \text{otherwise} \end{cases}$$

Here, $||g_i - p_i||$ represents the Euclidean distance between the ground truth keypoint $g_i$ and the predicted keypoint $p_i$.

The PCK score is then calculated as the average of the indicator values over all keypoints. It represents the percentage of keypoints that have been correctly predicted within the specified threshold distance $\delta$. A higher PCK score indicates better accuracy and alignment between the predicted and ground truth keypoints.

In the context of sign language generation models, PCK can be used to evaluate the quality of the generated sign language poses by comparing them to the ground truth poses. However, it's important to note that PCK only considers individual keypoints and does not capture the overall spatial or temporal coherence of the generated sign language sequences.

### 4.1.4 Results and insights

We report DTW [14] and Probability of Correct Keypoints scores on the Indian Sign Language dataset and compare it with the results of both Text2Sign [32] and Speech2Sign [15] methods. From table 4.1 we observe that our model performs significantly better than the existing Speech2Sign [15] method. Figure 4.1 shows the sample qualitative results. An interesting observation from the provided sample results, as well as other instances in our evaluation, is that while our model encounters challenges in accurately capturing the precise positions of hands and facial features in specific frames, these representations exhibit a visual similarity to the target RGB frames. It is worth noting, however, that minor disparities in hand positions and facial expressions can convey substantially different meanings in sign language. Consequently, we refrain from drawing definitive conclusions from our qualitative assessments and defer such considerations to future research endeavors.

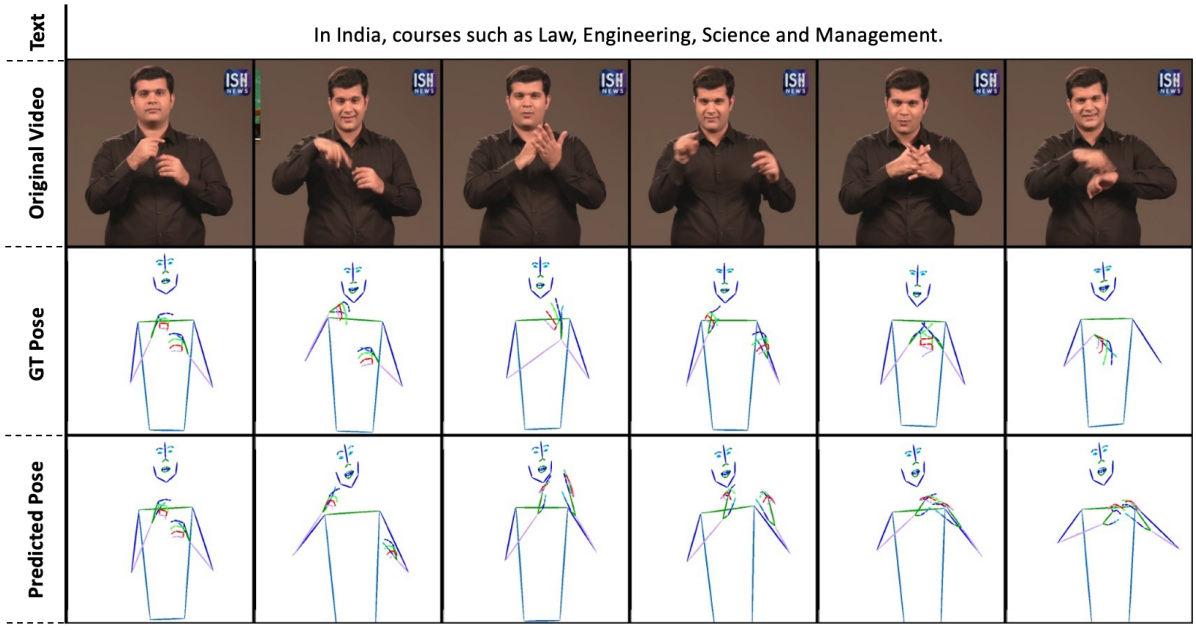


Figure 4.1: Qualitative Results illustrating the input text, the original video, the ground truth pose, and the predicted pose.

| Model | DTW Score ↓ | PCK ↑ |
|---|---|---|
| Dev set | | |
| Text ->Sign [32] | 19.55 | 0.61 |
| Speech2sign [15] | 15.94 | 0.72 |
| PE + TE ->Sign | 16.1 | 0.74 |
| PE + TE ->Sign + FAUs | **13.37** | **0.79** |
| Test set | | |
| Text ->Sign | 22.55 | 0.59 |
| Speech2sign [15] | 14.08 | 0.78 |
| PE + TE ->Sign | 17.3 | 0.72 |
| PE + TE ->Sign + FAUs | **13.37** | **0.79** |

Table 4.1: Comparison of Dynamic Time Warping (DTW) and Probabilty of Correct Keypoints (PCK) scores with baselines on dev and test sets. B+F indicates model that predicts body+face keypoints. PE - Prosody Encoder; TE: Text Encoder

### 4.1.5 Ablation Analysis

To evaluate the contribution of each component in our proposed architecture, we conduct ablation studies on our model. Specifically, we perform experiments where we remove each component from the multitasking setup one by one and compare the results with the full model.

Table 4.2 summarizes the results of our ablation studies. As can be seen, removing the FAUs decoder results in a drop in performance in both metrics. The results demonstrate the effectiveness of our multitasking approach in leveraging multiple modalities for sign language generation.

We observe that the results of our final model are still close to the model that uses just the text encoder to predict only sign-pose sequences. However, when trying to predict both sign-poses and FAUs, the same approach suffers in comparison. We want to highlight that even though the sign poses include facial keypoints, their range of movements is limited when compared to hands and other body keypoints thus limiting their contribution in the supervision of the models. The FAUs capture facial expressions more holistically and their prediction, posed as a multi-label classification task, adds to the difficulty of the proposed task. Additionally, we note that the same task of predicting FAUs helps boost the performance of the prosody-encoder only model as they likely provide better supervision for aligning the prosodic element inputs with candidate sign-pose sequences in the latent space. In

summary, our ablation studies demonstrate the effectiveness of our multitasking approach in leveraging multiple modalities for sign language generation.

| Model | DTW Score ↓ | PCK ↑ |
|---|---|---|
| TE ->Sign | 13.82 | 0.81 |
| TE ->Sign + FAUs | 15.69 | 0.78 |
| PE ->Sign | 17.16 | 0.73 |
| PE ->Sign + FAUs | 14.52 | 0.75 |
| PE + TE ->Sign + FAUs (Ours) | **13.23** | **0.81** |

Table 4.2: Comparison of ablation studies. PE - Prosody Encoder; TE-Text Encoder

### 4.1.6 Implementation Details

We set up our transformer model with two layers for both encoders and decoders, each equipped with eight attention heads. Both encoders and decoders use a hidden size of $512$. We use the Adam optimiser with an initial learning rate of $0.001$, which can be reduced if the training plateaus. We apply gradient clipping with a threshold of $5.0$ and use a batch size of $32$ for training efficiency. We incorporate Future Prediction as proposed by [32]. The training loss function includes L1 regularisation along with losses for specific components, each weighted accordingly. For the loss function, the values for $\lambda_{\text{Pose}}, \lambda_{\text{FAUs}}, \lambda_{\text{Discriminator}}$ are $1$, $0.001$, $0.0001$ respectively.

*Chapter 5*

# Challenges and Limitations

In this chapter, we delve into the challenges and limitations encountered in developing sign language generation models. Understanding and addressing these challenges is crucial for refining the model. For example, we highlight the importance of human evaluation by sign language experts to ensure real-world effectiveness. Additionally, we acknowledge the model's difficulty in representing subtle movements and the need to consider factors beyond hand and facial expressions. This chapter sheds light on the complexities of extracting accurate poses and urges exploration into alternative representations. It also highlights the impact of dataset limitations and individual signer styles, providing valuable insights for future improvements.

**Evaluation Methods:** Although our model has achieved state-of-the-art results based on DTW scores, it is essential to conduct human evaluation with expert sign language interpreters to ensure the quality and relevance of the generated sign language. DTW scores only assess the alignment between ground truth poses and predicted poses but do not measure the correlation with the input speech. Correlating these scores with human evaluation ratings is crucial for understanding the model's performance in real-world communication scenarios. Metrics that measure the coherence and synchronization of other non-manual elements, such as body posture, head movements, and eye gaze are also necessary [39]. Therefore, when designing a sign language generation model, accounting for these linguistic elements and their dynamic interactions is essential to produce more accurate and culturally appropriate sign language outputs.

**Fine Movements:** The current model successfully learns coarse hand movements but lacks the ability to capture fine movements of fingers and facial parts (See Figure 5.1). This limitation is attributed to the use of Mean Squared Error (MSE) loss, which penalizes larger movements more than fine movements. To address this issue, alternative loss functions, such as a keypoint loss proposed by [34], should be explored. This loss involves a hand keypoint discriminator pre-trained on 2D hand poses and may improve the model's capability to generate more accurate and intricate hand movements.

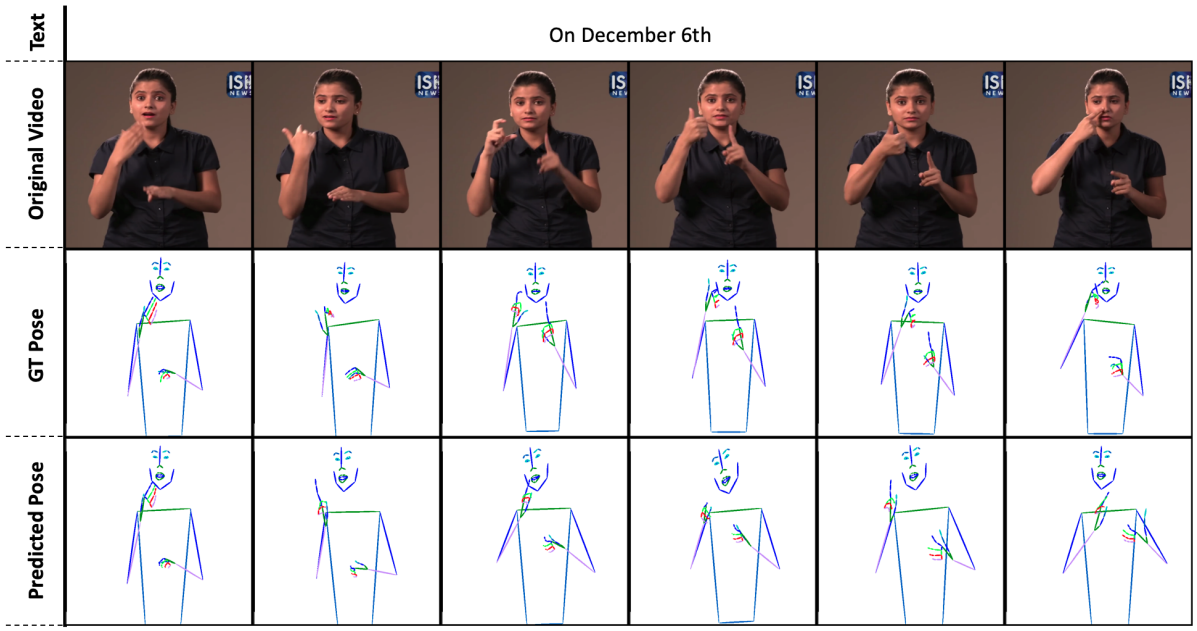


Figure 5.1: Sample result showing the model's accurate hand movement prediction with inaccurate finger movements.

**More Linguistic Information:** One significant challenge lies in handling the sequential nature of input speech or text, as opposed to the non-linear and simultaneous nature of sign language. Speech unfolds in a linear manner, and sign language relies on the integration of multiple components in parallel. Thus, capturing and mapping these linguistic structures effectively requires specialized attention. Understanding how signers use space, directionality, and facial expressions to indicate different grammatical constructs is crucial for generating natural and contextually appropriate sign language. Currently, our model focuses primarily on generating hand and facial movements, neglecting other crucial components. Future work should explore incorporating non-manual markers, body language, and gaze direction into the generation process to enhance the naturalness and comprehensiveness of sign language communication.

**Errors in Skeleton Pose Extraction:** One of the significant challenges in sign language generation is accurately extracting the skeleton pose from the input video or speech. The skeleton pose serves as a crucial input to the model, representing the keypoint positions of the signer's hands, face, and body movements. Although advanced pose estimation techniques like Mediapipe provide robust keypoint predictions, there are inherent limitations and errors that can impact the overall performance of the sign language generation model. Sign language videos captured in real-world settings may contain various forms of noise, occlusions, and artifacts. These imperfections can lead to inaccuracies in the pose estimation process, resulting in incorrect keypoint positions. For instance, background clutter, complex hand gestures, or fast movements may obscure the hand keypoints, leading to incomplete or noisy pose

representations. Figure 5.2 shows Additionally, sign language involves intricate hand and finger movements that can sometimes be challenging to discern accurately. The dynamic nature of sign language requires precise identification of hand shapes, finger positions, and gestures. However, the inherent ambiguity in certain signs or gestures can lead to misinterpretations and inaccuracies in the extracted skeleton pose.

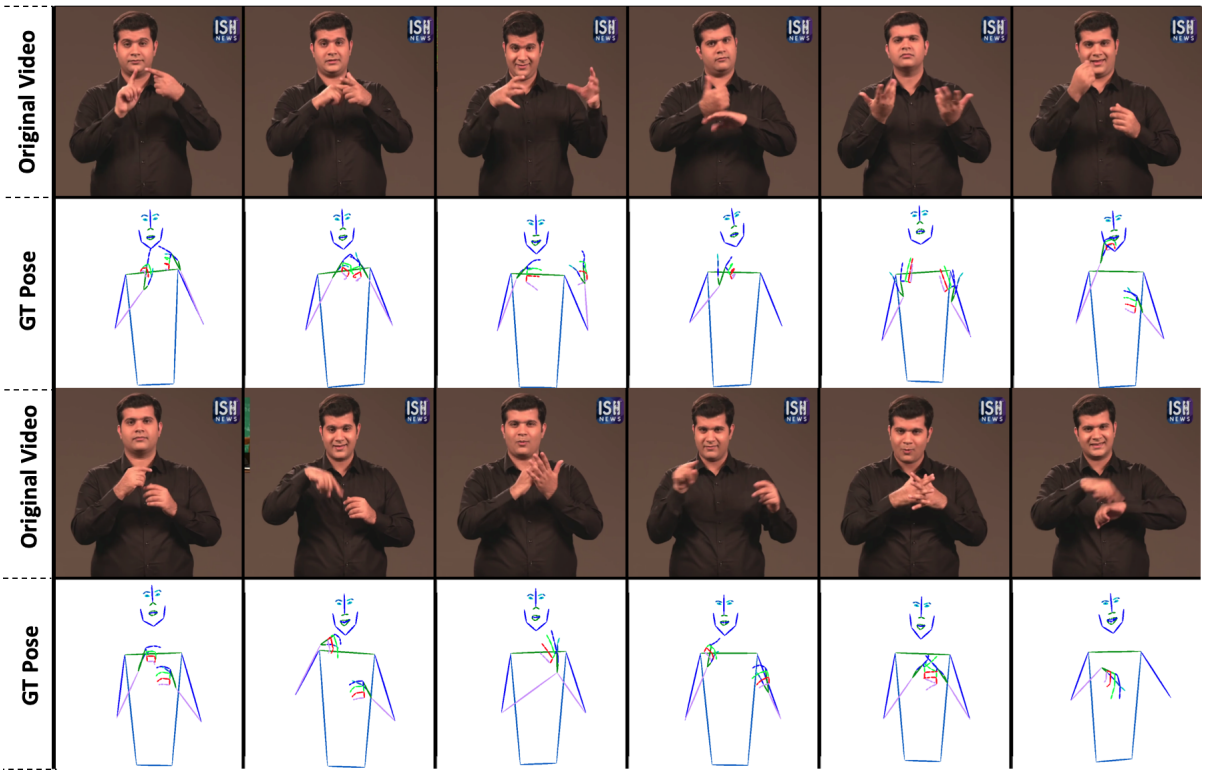


Figure 5.2: Mediapipe Errors. The keypoints for the fourth frame in the first video and the sixth frame in the second video are predicted incorrectly due to fast/blurry movements whereas the keypoints for the third frame in the second video are predicted incorrectly as it contains a complex hand gesture.

**Pose Representation:** The representation of sign language as keypoint sequences in videos is abstract and results in the loss of some skeletal information. This may lead to some loss of fine-grained details in the generated sign language. Future research could explore alternative representations that preserve more intricate skeletal information for more accurate sign language generation.

**Dataset Size and Variety:** Our current dataset size and variety might be limited, which could impact the model's ability to capture the full complexity and richness of sign language. Expanding the dataset or exploring low-resource training techniques is essential to improve the model's generalization

and performance on diverse signing styles and linguistic patterns.

**Signer Style:** Sign language relies on the signer's individual style and preferences, which can significantly affect the model's performance. Investigating the impact of varying signer styles on the model's output and devising methods to adapt the model to different signing styles are critical for real-world applicability.

In conclusion, while our model shows promising results in generating sign language from speech, there are several limitations and challenges that need to be addressed in future work. This chapter offers essential reflections on the model's limitations, guiding future efforts to enhance its accuracy and effectiveness in diverse communication scenarios.

*Chapter 6*

# Conclusions and Future work

## 6.1 Ethical Considerations in Assistive Technology

In our study, it is important to acknowledge that we have employed a limited dataset of Indian sign language videos, primarily sourced from YouTube. While this dataset served as a valuable starting point for our investigation into speech-to-sign language generation models, we recognise its inherent limitations regarding representativeness for the broader sign language community. Machine learning models, while potent tools, are not immune to perpetuating societal biases embedded within their training data. As the model learns from diverse inputs, there is an inherent risk of it inadvertently internalizing existing biases, which can manifest in the generated sign language poses. Therefore, a critical ethical consideration lies in systematically identifying and rectifying biases to ensure the generated communication remains free from discrimination or misrepresentation. It is essential to emphasize that the models proposed in this paper are only to explore the role of prosody in speech-sign language generation models and are not suitable for direct deployment due to their insufficient scope and potential biases. Moreover, we acknowledge that a critical aspect, validation with signers, has not been fully undertaken within the scope of this study. This is a significant limitation that warrants further attention and validation in future research endeavours.

## 6.2 Conclusion

In this thesis, we first discuss the complexity of sign language and the the essential role of non-manual elements in sign language, including their functions in grammar, pragmatics, and discourse. Furthermore, we delved into the significance of recognizing non-manual expressions in the context of sign language processing, shedding light on the influence of prosodic information in spoken language and the pivotal role of Facial Action Units (FAUs) in conveying sign language expressions.

In Chapter 2, we begin with the discussionn of early methods in sign language generation primarily relying on text or gloss inputs. These methods fall short in capturing the richness of sign language, notably facial expressions and body language. Despite attempts to include both manual and

non-manual features, a significant reliance on text persists, highlighting the need for more comprehensive approaches. The chapter then delves into the critical role of non-manual markers, emphasizing their functions in grammar, pragmatics, and discourse. The challenges associated with annotating the facial expressions and utilizing pre-trained models for weak labels were also highlighted. The discussion extends to co-speech gesture generation, underscoring the importance of combining audio and text modalities. This chapter lays the groundwork for our approach by identifying gaps and opportunities in current sign language generation methods, paving the way for a more holistic and effective model.

Building on this theoretical foundation, we introduce our approach to sign language generation in Chapter 3, which aims to generate sign language poses from input speech and text. Employing a multi-task learning strategy, our architecture integrates a speech encoder, a Facial Action Units (FAUs) decoder, and a sign pose decoder. The overall model, illustrated in Figure 3.1, relies on BERT embeddings for text and Tacotron 2 GST encodings for audio to extract semantic and prosodic information. We emphasize the significance of FAUs in capturing facial expressions and introduce a preprocessing pipeline using the ME-GraphAU model for weak supervision. The FAUs prediction is integrated as an additional task, enhancing the model's understanding of facial cues. The key model components include a prosody encoder, FAUs decoder, sign pose decoder, and a Cross-Modal Discriminator. The latter evaluates the alignment between speech and sign pose sequences. The multi-tasking setup facilitates joint learning of facial expressions, prosody, and sign pose generation, with a weighted sum of losses ensuring a balanced optimization approach. This comprehensive methodology lays the foundation for our model's ability to generate contextually rich and expressive sign language sequences.

In Chapter 4, we conduct a thorough performance evaluation of our model, encompassing various aspects such as dataset characteristics, baseline models, evaluation metrics, results, ablation analysis, and implementation details. Our model is compared against two baseline approaches, Text2Sign [32] and Speech2Sign [15], with the evaluation metrics including Dynamic Time Warping (DTW) and Probability of Correct Keypoints (PCK). The results unequivocally highlight the superiority of our proposed model, particularly when incorporating Facial Action Units (FAUs) prediction, showcasing its efficacy in generating accurate and context-aware sign pose sequences. Ablation studies further underscore the importance of the multitasking setup, emphasizing the effectiveness of harnessing multiple modalities for sign language generation. The implementation details include the model architecture, optimization parameters, and loss functions, contributing to a comprehensive understanding of our experimental framework.

In Chapter 5, we delve into the limitations and challenges of our sign language generation model. Firstly, we highlight the need for human evaluation involving sign language experts to ensure the contextual and qualitative relevance of generated sign language. This is especially crucial in assessing the model's real-world communication effectiveness, considering DTW scores focus solely on pose alignment and do not capture correlations with input speech. Additionally, the model's proficiency in capturing fine movements of fingers and facial parts is hindered by the Mean Squared Error (MSE) loss, necessitating exploration of alternative loss functions to enhance accuracy in intricate hand movements.

Despite achieving state-of-the-art results, our model's exclusive focus on hand and facial movements neglects other non-manual markers such as body language, and gaze direction, urging future exploration into a more comprehensive representation of sign language communication. Furthermore, challenges persist in accurate skeleton pose extraction due to real-world video complexities, noise, and occlusions, impacting the precision of keypoint predictions. The abstract nature of keypoint sequence representation poses a risk of losing fine-grained details in sign language, motivating future research into alternative representations. Issues related to dataset size, variety, and signer styles underscore the need for expanded datasets, low-resource training techniques, and adaptive modeling strategies for improved generalization and real-world applicability.

As we conclude this thesis, we recognize that while our approach has demonstrated promise, challenges and limitations persist. We aspire for our work to serve as a catalyst, inspiring future research endeavors aimed at enhancing accessibility and inclusivity for the deaf and hard-of-hearing community. Our thesis underscores the ongoing importance of exploring innovative solutions in the realm of assistive technology to foster improved communication and social inclusion for all.

# Related Publications

## 6.3 Relevant Publications

1. **Mounika Kanakanti**, Shantanu Singh, and Manish Shrivastava, *"MultiFacet: A Multi-Tasking Framework for Speech-to-Sign Language Generation,"* **INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23 Companion), October 9–13, 2023, Paris, France**

# Bibliography


[1] C. Ahuja, D. W. Lee, R. Ishii, and L.-P. Morency. No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1884–1895, Online, Nov. 2020. Association for Computational Linguistics.

[2] S. Alexanderson, G. E. Henter, T. Kucherenko, and J. Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum*, 39(2):487–496, 2020.

[3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[4] Blender Foundation. Blender, 2023. Computer software.

[5] N. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden. Neural sign language translation. 03 2018.

[6] M. Crawshaw. Multi-task learning with deep neural networks: A survey, 2020.

[7] E. P. da Silva, P. D. P. Costa, K. M. O. Kumada, and J. M. de Martino. Facial action unit detection methodology with application in brazilian sign language recognition. *Pattern Analysis and Applications*, 25:549 – 565, 2021.

[8] E. P. da Silva, K. M. O. Kumada, and P. D. P. Costa. Analysis of facial expressions in brazilian sign language (libras). *European Scientific Journal, ESJ*, 2021.

[9] S. Dachkovsky and W. Sandler. Visual intonation in the prosody of a sign language. *Language and Speech*, 52(2-3):287–314, 2009. PMID: 19624033.

[10] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[11] Ebling. SMILE Swiss German sign language dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[12] I. Grishchenko and V. Bazarevsky. Mediapipe holistic — simultaneous face, hand and pose prediction, on device, Dec. 2020.

[13] C. Gussenhhoven and A. Chen. *The Oxford Handbook of Language Prosody*. Oxford University Press, Dec. 2020.

[14] P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964.

[15] P. Kapoor, R. Mukhopadhyay, S. Hegde, V. Namboodiri, and C. Jawahar. Towards automatic speech to sign language generation. pages 3700–3704, 08 2021.

[16] O. Koller, J. Forster, and H. Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, Dec. 2015.

[17] T. Kucherenko, P. Jonell, S. van Waveren, G. E. Henter, S. Alexandersson, I. Leite, and H. Kjellström. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, page 242–250, New York, NY, USA, 2020. Association for Computing Machinery.

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

[19] C. Luo, S. Song, W. Xie, L. Shen, and H. Gunes. Learning multi-dimensional edge feature-based AU relation graph for facial action unit recognition. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, jul 2022.

[20] D. M. Madhiarasan and P. P. P. Roy. A comprehensive review of sign language recognition: Different types, modalities, and datasets, 2022.

[21] M. Mirza and S. Osindero. Conditional generative adversarial nets, 2014.

[22] M. Mukushev, A. Sabyrov, A. Imashev, K. Koishybay, V. Kimmelman, and A. Sandygulova. Evaluation of manual and non-manual components for sign language recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6073–6078, Marseille, France, May 2020. European Language Resources Association.

[23] G. NC, M. Ladi, S. Negi, P. Selvaraj, P. Kumar, and M. M. Khapra. Addressing resource scarcity across sign languages with multilingual pretraining and unified-vocabulary datasets. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[24] C. Neidle, A. Opoku, and D. N. Metaxas. ASL video corpora & sign bank: Resources available through the american sign language linguistic research project (ASLLRP). *CoRR*, abs/2201.07899, 2022.

[25] S. Nyatsanga, T. Kucherenko, C. Ahuja, G. E. Henter, and M. Neff. A comprehensive review of data-driven co-speech gesture generation. *CoRR*, abs/2301.05339, 2023.

[26] W. H. Organization. Hearing loss, 2023. Accessed: 21-07-2023.

[27] R. Pfau and J. Quer. *Nonmanuals: their grammatical and prosodic roles*, page 381–402. Cambridge Language Surveys. Cambridge University Press, 2010.

[28] R. Rastgoo, K. Kiani, S. Escalera, V. Athitsos, and M. Sabokrou. All You Need In Sign Language Production, Jan. 2022. arXiv:2201.01609 [cs].

[29] S. Ruder. An overview of multi-task learning in deep neural networks, 2017.

[30] B. Saunders, N. C. Camgoz, and R. Bowden. Adversarial training for multi-channel sign language production, 2020.

[31] B. Saunders, N. C. Camgoz, and R. Bowden. Everybody sign now: Translating spoken language to photo realistic sign language video, 2020.

[32] B. Saunders, N. C. Camgoz, and R. Bowden. Progressive Transformers for End-to-End Sign Language Production, July 2020. arXiv:2004.14874 [cs].

[33] B. Saunders, N. C. Camgoz, and R. Bowden. Skeletal graph self-attention: Embedding a skeleton inductive bias into sign language production, 2021.

[34] B. Saunders, N. C. Camgoz, and R. Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production, 2022.

[35] E. P. d. Silva, P. D. P. Costa, K. M. O. Kumada, and J. M. De Martino. Silfa: Sign language facial action database for the development of assistive technologies for the deaf. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 688–692, 2020.

[36] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4645–4653, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society.

[37] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden. Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. *Int. J. Comput. Vision*, 128(4):891–908, apr 2020.

[38] M. Theune, E. Klabbers, J. D. Pijper, E. Krahmer, and J. Odijk. From data to speech: a general approach. *Natural Language Engineering*, 7:47 – 86, 2001.

[39] E. Vahdani, L. Jing, Y. Tian, and M. Huenerfauth. Recognizing American Sign Language Nonmanual Signal Grammar Errors in Continuous Videos, May 2020. arXiv:2005.00253 [cs].

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need, Dec. 2017. arXiv:1706.03762 [cs].

[41] C. Viegas, M. İnan, L. Quandt, and M. Alikhani. Including Facial Expressions in Contextual Embeddings for Sign Language Generation, Feb. 2022. arXiv:2202.05383 [cs].

[42] H. Walsh, B. Saunders, and R. Bowden. Changing the representation: Examining language representation for neural sign language production. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 117–124, Marseille, France, June 2022. European Language Resources Association.

[43] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis, 2018.

[44] P. Wolfert. Genea challenge 2022, 2022. https://genea-workshop.github.io/2022/challenge/#home.

[45] K. Yin, A. Moryossef, J. Hochgesang, Y. Goldberg, and M. Alikhani. Including Signed Languages in Natural Language Processing. *arXiv:2105.05222 [cs]*, July 2021. arXiv: 2105.05222.

[46] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, and G. Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Trans. Graph.*, 39(6), nov 2020.

[47] J. Zelinka and J. Kanis. Neural sign language synthesis: Words are our glosses. pages 3384–3392, 03 2020.

[48] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu. Libritts: A corpus derived from librispeech for text-to-speech, 2019.

[49] C. Zhou, T. Bian, and K. Chen. Gesturemaster: Graph-based speech-driven gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, page 764–770, New York, NY, USA, 2022. Association for Computing Machinery.

[50] M. İnan, Y. Zhong, S. Hassan, L. Quandt, and M. Alikhani. Modeling Intensification for Sign Language Generation: A Computational Approach. 2022. Publisher: arXiv Version Number: 1.