EPFL

Thèse n° 11178

# Advancing Phonology-Based Sign Language Assessment: From Learner to Machine-Generated Videos

Présentée le 1er septembre 2025

Faculté des sciences et techniques de l'ingénieur
Laboratoire de l'IDIAP
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

**Neha TARIGOPULA**

Acceptée sur proposition du jury

Prof. P. Frossard, président du jury
Dr J.-M. Odobez, Dr M. Magimai Doss, directeurs de thèse
Prof. G. Potamianos, rapporteur
Dr C. Lea, rapporteur
Prof. J.-Ph. Thiran, rapporteur

École polytechnique fédérale de Lausanne

2025

To my family, friends and mentors ...

# Acknowledgments

I started my PhD journey with the dream of scaling new heights. I knew there would be ups and downs, but little did I know that the downs would rush past like the wind in your hair as you swoosh down the slopes of Tsantonnaire: fast, slightly out of control, and full of unexpected lessons. The ups, on the other hand, felt more like climbing Pierre Avoi with your bare hands: slow, intense, and filled with moments of wondering why you ever thought this was a good idea. I am grateful to all the people who were a part of this alpine journey. Thank you for cheering me on, for picking me up, and guiding me to ensure I did not get lost along the way.

At the heart of this journey was my advisor, Dr. Mathew Magimai Doss, whose support and guidance were invaluable. His steady mentorship, intellectual clarity, and attention to detail have shaped not only this thesis but also my growth as a researcher. I am especially thankful for the thoughtful balance he maintained between offering direction and allowing me the intellectual freedom to explore my own ideas.

I also express my sincere thanks to my thesis director, Prof. Jean-Marc Odobez, for his suggestions and feedback throughout this process. I thank my jury members, Prof. Gerasimos Potamianos, Dr. Colin Lea, Prof. Jean-Philippe Thiran, and Prof. Pascal Frossard, for their valuable insights, thought-provoking questions, and generous engagement with my work. The administrative staff at Idiap have been supportive since the very beginning, and their help in easing my transition to life in Switzerland has been truly appreciated.

This research was made possible by funding from the Swiss National Science Foundation (SNSF) through the Sinergia project SMILE-II (Scalable Multimodal sign language technology for sIgn language Learning and assessmEnt -II), grant no. CRSII5_193686. I am grateful for the funding that supported my research and enabled me to present my work at various conferences. Through this project, I had the privilege of collaborating with a diverse group of researchers from Hochschule für Heilpädagogik (HfH), the University of Zurich, and the University of Surrey. These collaborations enriched my research experience and broadened my perspective in meaningful ways. I would like to extend special thanks to Prof. Richard Bowden for hosting me at the CVSSP lab at the University of Surrey for two months. Working alongside his team was a truly enriching experience, I learned a great deal and deeply appreciated the open, collaborative environment. I am also thankful to Sandrine Tornay and Skanda Muralidhar for being wonderful colleagues and for their supportive presence throughout our

collaboration. I am especially grateful to Sandrine, whose kindness and encouragement have meant a great deal to me.

My alpine journey, true to its name, was a product of being in the company of the mountains I saw through my window. They filled my heart, and if I thought there was room left, I filled it up with some raclette. One of the best parts of this journey at Idiap was the people I met along the way, each unique in their own way. I am fortunate to have met Amir, Angel, Andrei, Anshul, Carlos, Chloé, Darya, Enno, Fabius, Florian, François, Julian, Karl, Laura, Laurent, Louise, Mirko, Olena, Pablo, Pierre, Roberto, Sarthak, Shashi, Sheiklavya, Simon, Suhan, Valentin, Weipeng, Zohreh, and many others that introduced a part of their world to me. It was lovely sharing a home with Pablo and Suhan, they have evolved to become constant sources of support, and for that I'm grateful. I would especially like to thank Florian for being a wonderful friend, for sharing his love of the mountains, and for always being there to share a laugh or a meal.

I express my gratitude to my first mentor, Prof. Dinesh Jayagopi at IIIT Bangalore, who introduced me to the world of research and encouraged me at every step of my career. I also thank the Neha who bravely left Mercedes Research to chase this dream. A huge shoutout to Saikumar Dwivedi, my mentor and friend, without whom I would not be here today. His constant motivation and support truly paved the way for my PhD journey.

I'm extremely lucky to have friends I have known for half my life. Anusha, Anushka, Laasya, Manasa, Sahiti, and Valli, thank you for being my pillars of support and for sharing this journey with me, even from afar. Laasya, thank you for being at the receiving end of my daily roller coasters and for always enabling my crazy.

I am deeply grateful to my family for their unwavering support and love, especially my parents, Rajashekar and Jagadeshwari, and my brother, Sanjeev. I'm thankful to my parents for standing by every decision I've made and for always believing in me, even when the path wasn't clear. Growing up, I tried to secretly follow in my brother's footsteps in countless ways, and the idea of doing a PhD was one of the many seeds he unknowingly planted. I also feel lucky to have grandparents who are endlessly proud and constantly cheering me on. Finally, I would like to thank my partner, Rohan, for quietly standing by my side and watching me grow over the years. Your patience, understanding, and love have been my anchor, and I'm truly grateful to have you in my life.

*Martigny, August 2, 2025* Neha

# Abstract

Sign languages are rich visual languages that use both manual features like handshape, movement, and location and non-manual features such as facial expressions, head movements, and body posture to convey information. Sign language learning technologies have the potential to bridge the communication gap between hearing and hard-of-hearing communities by providing accessible platforms that deliver actionable, interpretable feedback on signing performance. In the light of advancements in deep learning, sign language recognition has seen a plethora of improvements; however, sign language assessment, the task of evaluating the quality and correctness of signing, remains in its infancy. This thesis aims to advance assessment systems for sign language that not only provide feedback to help learners improve their signing performance, but also hold potential for offering automated feedback in the context of sign language generation. We build upon an explainable phonology-based framework, advancing it to work with RGB video input with the goal of enabling a webcam-based sign language assessment system that is accessible and suitable for real-world deployment. In this context, we investigate various 2D and 3D skeleton estimation methods derived from RGB input to extract interpretable and linguistically meaningful manual features. To incorporate more holistic representations, we further explore deep learning-based methods for modeling hand movement. Specifically, we evaluate spatio-temporal features from convolutional networks and sign language-agnostic Vision Transformers for assessing hand movement and handshape quality in isolated signs. We then investigate how assessment can be effectively leveraged for evaluating generated sign productions. To this end, we propose a posterior-based assessment framework that uses articulatory posteriors to evaluate the quality of generated signing. This approach enables interpretable, frame-level feedback and is applicable to both video-to-video and text-to-pose generation models. Finally, we examine the role of non-manual features such as facial expressions in continuous sign language, where they play a critical role in marking grammatical structure and sign boundaries. We develop Facial Action Unit-based non-manual feature detectors, integrate the resulting posteriors into our phonology-based framework, and then analyze their contribution to sign segmentation through alignment.

**Keywords**: Sign language learning, sign language generation, phonology-based sign language, sign language subunits, spatio-temporal features, posterior-based assessment, sign segmentation

# Résumé

Les langues des signes sont des langages visuels riches qui utilisent pour transmettre de l'information à la fois des éléments manuels tels que la forme de la main, son mouvement et sa position, et des caractéristiques non manuelles telles que les expressions faciales, les mouvements de tête et la posture du corps. Les technologies d'apprentissage de la langue des signes ont le potentiel de combler le fossé de communication entre les communautés d'entendants et de malentendants en offrant des outils accessibles qui fournissent un retour concret et interprétable sur la maîtrise de la langue des signes. Grâce aux progrès en apprentissage profond, la reconnaissance de la langue des signes (Sign Language Recognition, SLR) a connu une multitude d'améliorations ; cependant l'évaluation de la langue des signes (Sign Language Assessment, SLA), qui consiste à évaluer la qualité et la justesse des signes, en est encore à ses balbutiements. Cette thèse a pour but de faire progresser les systèmes d'évaluation de la langue des signes qui non seulement offrent un retour pour aider les apprenants à améliorer la qualité de leurs signes, mais qui ont aussi le potentiel de fournir une mesure de qualité automatisée dans le contexte de la génération de la langue des signes (Sign Language Generation, SLG). Nous nous basons sur un cadre explicable basé sur la phonologie et le développons pour traiter des vidéos RVB en entrée, afin de proposer un système d'évaluation de la langue des signes utilisant une webcam, qui est à la fois accessible et adapté pour un déploiement dans une situation réelle. Dans ce contexte, nous étudions diverses méthodes d'estimation du squelette en 2D et 3D sur base de la vidéo RVB pour extraire des caractéristiques manuelles interprétables et pertinentes au niveau linguistique. Afin d'intégrer des représentations plus globales, nous explorons plus avant les méthodes de modélisation du mouvement de la main basées sur l'apprentissage profond. Plus précisément, nous évaluons des caractéristiques spatio-temporelles afin de déterminer la forme et le mouvement de la main dans des signes isolés, à travers des réseaux de neurones convolutionnels entraînés sur des données de la langue des signes et des Vision Transformers qui n'ont pas été exposés à ce type de données.

Nous étudions ensuite comment la SLA peut être exploitée efficacement pour évaluer la SLG. Dans ce but, nous proposons un cadre d'évaluation basé sur les probabilités a posteriori qui utilise les probabilités des articulations pour mesurer la qualité des signes générés. Cette approche permet un retour interprétable image par image, et est applicable à la fois aux modèles de génération vidéo-à-vidéo et texte-vers-pose. Enfin, nous étudions le rôle des caractéristiques non manuelles telles que les expressions faciales dans la langue des signes

continue, où celles-ci sont déterminantes pour marquer la structure grammaticale et les limites des signes. Nous développons des détecteurs de caractéristiques non manuelles basés sur les unités d'action faciale (Facial Action Units), intégrons les probabilités résultantes dans notre cadre phonologique, puis analysons leur contribution à la segmentation des signes par alignement.

**Mots-clés** : Apprentissage de la langue des signes, génération de la langue des signes, langue des signes basée sur la phonologie, sous-unités de la langue des signes, caractéristiques spatio-temporelles, évaluation basée sur les probabilités a posteriori, segmentation des signes.

# Zusammenfassung

Gebärdensprachen sind reichhaltige visuelle Sprachen, die zur Übermittlung von Informationen sowohl manuelle Elemente wie Handform, -bewegung und -position als auch nichtmanuelle Merkmale wie Mimik, Kopfbewegungen und Körperhaltung nutzen. Technologien zum Erlernen der Gebärdensprache haben das Potenzial, die Kommunikationslücke zwischen hörenden und hörgeschädigten Gemeinschaften zu schließen, indem sie zugängliche Tools bieten, die konkretes und interpretierbares Feedback zur Beherrschung der Gebärdensprache liefern. Dank der Fortschritte im Bereich des Deep Learning hat die Erkennung von Gebärdensprache (Sign Language Recognition, SLR) eine Vielzahl von Verbesserungen erlebt; die Bewertung von Gebärdensprache (Sign Language Assessment, SLA), bei der die Qualität und Richtigkeit der Gebärden bewertet wird, steckt jedoch noch in den Kinderschuhen. Das Ziel dieser Arbeit ist es, Gebärdensprache-Bewertungssysteme weiterzuentwickeln, die nicht nur Feedback geben, um Lernenden zu helfen, die Qualität ihrer Gebärden zu verbessern, sondern auch das Potenzial haben, eine automatisierte Qualitätsmessung im Kontext der Gebärdensprache-Generierung (Sign Language Generation, SLG) zu liefern. Wir stützen uns auf einen erklärbaren, auf Phonologie basierenden Rahmen und entwickeln ihn weiter, um RGB-Videos als Eingabe zu verarbeiten, um ein Gebärdensprache-Bewertungssystem unter Verwendung einer Webcam anzubieten, das sowohl zugänglich als auch für den Einsatz in einer realen Situation geeignet ist. In diesem Zusammenhang untersuchen wir verschiedene Methoden zur 2D- und 3D-Skelettschätzung auf der Grundlage von RGB-Videos, um interpretierbare und sprachlich relevante Handmerkmale zu extrahieren. Um umfassendere Darstellungen zu integrieren, untersuchen wir weiter Methoden zur Modellierung von Handbewegungen auf der Grundlage von Deep Learning.

Genauer gesagt bewerten wir räumlich-zeitliche Merkmale, um die Form und Bewegung der Hand in einzelnen Gebärden zu bestimmen, und zwar mithilfe von konvolutionellen neuronalen Netzen, die mit Gebärdensprachdaten trainiert wurden, und Vision Transformers, die dieser Art von Daten nicht ausgesetzt waren.

Anschließend untersuchen wir, wie SLA effektiv zur Bewertung von SLG genutzt werden kann. Zu diesem Zweck schlagen wir einen auf A-posteriori-Wahrscheinlichkeiten basierenden Bewertungsrahmen vor, der die Wahrscheinlichkeiten der Gelenke nutzt, um die Qualität der generierten Gebärden zu messen. Dieser Ansatz ermöglicht ein interpretierbares Feedback Bild für Bild und ist sowohl auf Video-zu-Video- als auch auf Text-zu-Pose-Generierungsmodelle

anwendbar. Schließlich untersuchen wir die Rolle nicht-manueller Merkmale wie Mimik in der kontinuierlichen Gebärdensprache, wo diese entscheidend sind, um die grammatikalische Struktur und die Grenzen der Gebärden zu markieren. Wir entwickeln Detektoren für nicht-manuelle Merkmale auf der Grundlage von Facial Action Units (Gesichtsaktions-Einheiten), integrieren die resultierenden Wahrscheinlichkeiten in unser phonologisches Framework und analysieren dann ihren Beitrag zur Segmentierung von Zeichen durch Ausrichtung.

**Schlüsselwörter**: Erlernen der Gebärdensprache, Gebärdensprache-Generierung, phonologiebasierte Gebärdensprache, Untereinheiten der Gebärdensprache, räumlich-zeitliche Merkmale, Bewertung auf Basis von A-posteriori-Wahrscheinlichkeiten, Segmentierung von Gebärden.

# Contents

# List of Figures

# List of Tables

# 1 Introduction

Sign languages are full-fledged natural languages that rely on a complex interplay of manual features such as handshape, movement, location, and orientation and non-manual markers including facial expressions and body posture. They are the primary means of communication for deaf and hard-of-hearing individuals worldwide and play a central role in Deaf culture and identity. Importantly, sign languages are not universal. Just like spoken languages, they have evolved independently within different Deaf communities, shaped by local histories, cultures, and social contexts. This linguistic diversity poses additional challenges for computational modeling, especially in creating generalizable tools that can respect and adapt to the nuances of specific sign languages. Unlike spoken languages, sign languages lack a standardized written form. The two most commonly used representations are:

1. *Gloss:* a gloss is a word from the local spoken language that approximates the meaning of a sign. It serves as a semantic label but does not capture the phonological form or visual articulation of the sign.

2. *HamNoSys*[1]: the Hamburg Notation System is a phonetic transcription system that encodes handshape, location, orientation, and movement using a set of abstract symbols.

Recent advances in sign language processing have enabled significant progress in automatic sign language recognition, allowing machines to classify isolated signs or continuous sequences with increasing accuracy. These developments have been largely driven by deep learning methods and the availability of annotated datasets. However, the field has yet to develop robust tools for automated sign language assessment: tools that go beyond recognition to evaluate how accurately and clearly a sign is produced, with the ability to explain *why* a sign may be incorrect.

[1] https://www.sign-lang.uni-hamburg.de/dgs-korpus/hamnosys-97.html

## 1.1 Motivation

While recognition systems have made substantial progress, sign language learners especially those without regular access to native signers or qualified instructors continue to face a major challenge: the lack of constructive feedback on their signing performance. In most current systems, learners are only told whether the sign produced was correct or not, without insight into why a sign may be incorrect. This is a fundamental limitation in the context of education, where targeted, interpretable feedback is essential for effective learning. Importantly, this type of interpretable feedback can also be extended to the automatic assessment of sign language proficiency tests. Figure 1.1 illustrates the overall sign language assessment framework designed to meet these needs.

Moreover, as generative models (Goodfellow et al., 2020; Blattmann et al., 2023) for sign language such as those that synthesize signing videos or motion sequences become more prevalent, the need for objective, automated evaluation becomes even more pressing. These models often produce outputs that are visually plausible but may lack linguistic accuracy or articulatory correctness. Without dedicated tools to assess the quality and correctness of these generated signs, there is a risk of propagating unrealistic or incorrect signing patterns, particularly when such outputs are used in learning or accessibility contexts.

*Together, these challenges highlight the need for automated sign language assessment methods that are not only accurate but also interpretable and applicable to both human and machine-generated signing.*

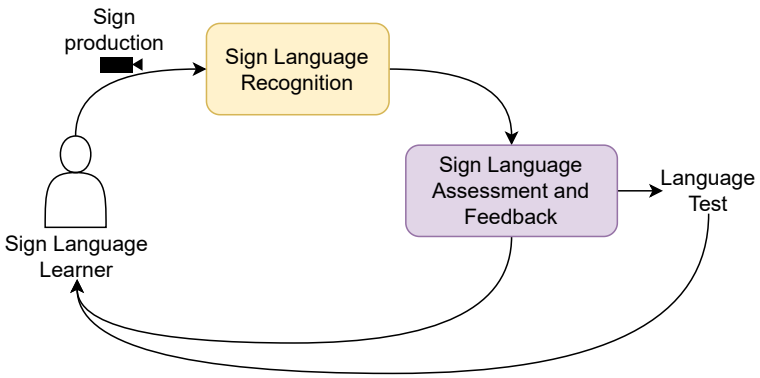


**Figure 1.1** – Sign language assessment framework. The framework is designed to provide interpretable feedback on sign language production. It integrates various components for sign language recognition, assessment, and feedback generation.

The work in this thesis is taking place in the context of Swiss National Science Foundation (SNSF) Sinergia SMILE-II (Scalable Multimodal sign language technology for sIgn language Learning and assessmEnt -II), a project that aims to build technology for sign language learning. More specifically, it builds on the groundwork laid by the project SMILE, described in Chapter 2, that aimed at recognition and assessment of isolated Swiss German sign language. SMILE-II aims to extend this technology to include both manual and non-manual features in

the context of continuous sign language recognition and sentence level-assessment. In this context, the SMILE-II project brings together researchers from different interdisciplinary fields of sign language linguistics, sign language assessment, computer vision, sign language technology, and spoken language technology, across four different institutions, namely, Idiap Research Institute (Martigny, Switzerland), HfH[2] Zurich, University of Zurich, and the University of Surrey (UK).

## 1.2 Contributions

The contributions of this thesis are as follows:

- **RGB-based sign language assessment methods for accessible sign language learning.**

  We explore different 2D and 3D skeleton estimation methods from RGB videos to be integrated with webcam-based systems for accessible sign language learning and assessment. We systematically evaluate the performance of these methods with the existing system and highlight the trade-off between sign language recognition performance and assessment performance.

  Parts of this work have appeared in:

  > Neha Tarigopula, Sandrine Tornay, Skanda Muralidhar, and Mathew Magimai Doss (2022). "Towards Accessible Sign Language Assessment and Learning". In: *Proceedings of the 2022 International Conference on Multimodal Interaction. ICMI '22. Bengaluru, India: Association for Computing Machinery,* pp. 626–631. DOI: 10.1145/3536221. 3556623

  > S. Tornay, A. Nanchen, A. Battisti, F. Holzknecht, N. Tarigopula, O. Mendez Maldonado, N. C. Camgöz, M. Razavi, K. Tissi, S. Sidler-Miserez, P. Boyes Bream, S. Ebling, T. Haug, R. Bowden, and M. Magimai-Doss (June 2023). "Web SMILE demo: a web application providing automated feedback on sign language vocabulary production". *In: 44th Language Testing and Research Colloquium: Language Assessment for a Global, Digital, and More Equitable Era.* Demo presentation. New York City, USA

- **Using deep-learning based methods for hand movement modeling in sign language assessment.**

  We employ deep learning models to extract spatio-temporal representations from RGB videos and evaluate their effectiveness in capturing articulatory quality, particularly hand movement. This includes the use of convolutional models like I3D as well as self-

[2] HfH stands for Hochschule für Heilpädagogik

supervised Vision Transformers (ViTs). We perform a layer-wise analysis of ViT features and highlight their strengths and limitations in terms of phonological interpretability.

Parts of this work have appeared in:

> Neha Tarigopula, Sandrine Tornay, Ozge Mercanoglu Sincan, Richard Bowden, and Mathew Magimai.-Doss (2025). "Posterior-Based Analysis of Spatio-Temporal Features for Sign Language Assessment". In: *IEEE Open Journal of Signal Processing 6*, pp. 284– 292. Presented at ICASSP 2025. DOI: 10.1109/OJSP.2025.3531781

- **Towards closing the feedback loop in sign language generation with assessment.**

  We propose a posterior-based framework for the fine-grained assessment of automatically generated sign language content. Using articulatory posteriors, we evaluate the quality of signs produced by video-to-video and text-to-pose generation models, focusing on hand movement and handshape. This provides interpretable analysis of articulatory accuracy and lays the groundwork for diagnostic evaluation in sign language generation.

  Parts of this work have appeared in:

  > Neha Tarigopula, Preyas Garg, Skanda Muralidhar, Sandrine Tornay, Dinesh Babu Jayagopi, and Mathew Magimai.-Doss (2024). "Content-Based Objective Evaluation of Artificially Generated Sign Language Videos". In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3815–3819. DOI: 10.1109/ICASSP48485.2024.10448192

  > Zifan Jiang, Colin Leong, Amit Moryossef, Anne Göhring, Annette Rios, Oliver Cory, Maksym Ivashechkin, Neha Tarigopula, Biao Zhang, Rico Sennrich, and Sarah Ebling. "Meaningful Pose-Based Sign Language Evaluation". *Manuscript under review, 2025*

- **Non-manual feature detection for continuous sign language.**

  We develop and compare non-manual feature detection methods, including spatio-temporal and transformer-based models, in the context of continuous sign language. These features are integrated as independent channels into a phonology-based alignment framework and evaluated on the SMILE-SRT dataset. We assess their contribution to sign segmentation, a critical first step toward linguistically grounded continuous sign language assessment.

## 1.3 Thesis outline

The remainder of this thesis is organized as follows: In Chapter 2, we provide a detailed background of sign language processing technologies, focusing on sign language recognition, assessment, and generation. We then discuss in detail the existing methods for sign language assessment and detail the datasets used in this thesis.

In Chapter 3, we present the work on accessible sign language assessment methods. We evaluate the methods using SMILE-DSGS dataset. In Chapter 4, we focus on the use of deep learning methods for hand movement modeling in sign language assessment. Building on the skeleton-based and RGB-based methodologies developed in Chapter 3 and Chapter 4, we propose a posterior-based approach to assess generated sign language content in Chapter 5. In Chapter 6, we compare different non-manual feature detection methods for continuous sign language and their integration into the phonological framework for sign alignment. Finally, in Chapter 7, we summarize this thesis and suggest directions for future work.

# 2 Background, Methods and Datasets

## 2.1 Background

Sign Language Processing (SLP) is an evolving field that encompasses various technologies and methodologies aimed at understanding, generating, and assessing sign languages. Our focus remains on SLP within the realm of linguistics and computer vision. This chapter provides an overview of the key components of SLP, including Sign Language Recognition (SLR) and Sign Language Translation (SLT), Sign Language Generation (SLG), and Sign Language Assessment (SLA). We will also discuss the significance of these technologies in enhancing communication accessibility for Deaf and hard-of-hearing individuals. Figure 2.1 illustrates the taxonomy of SLP technologies.

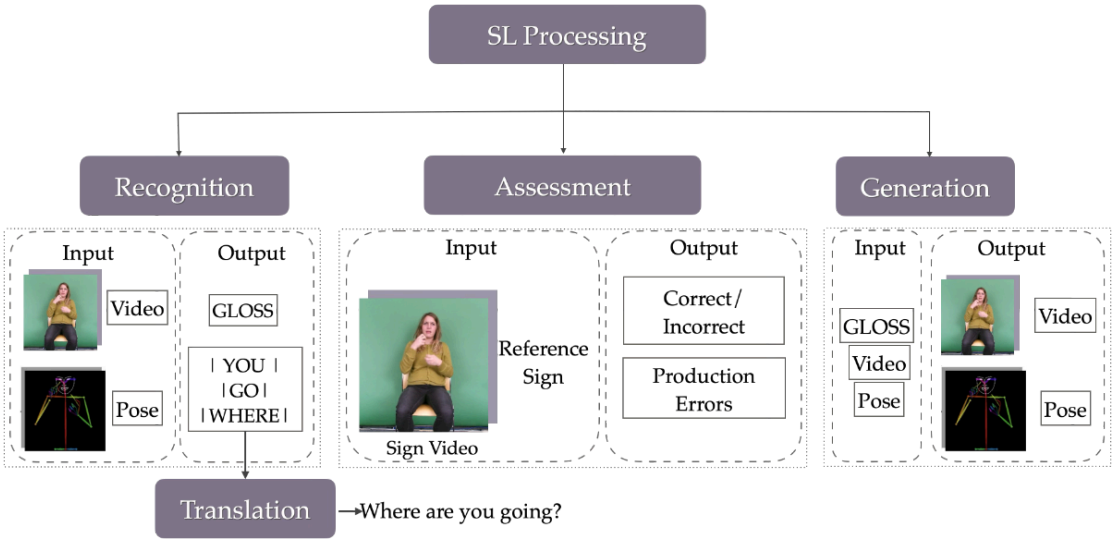


**Figure 2.1** – Taxonomy of sign language processing technologies.

### 2.1.1 Sign language recognition

SLR refers to the task of automatically recognizing signs from visual input. The goal is to map the visual input to glosses in a specific sign language. SLR can be categorized into three types:

1. Isolated SLR, where individual signs are recognized in segmented clips.
2. Continuous SLR, where a sequence of signs is recognized from unsegmented signing streams, often involving co-articulation and transitional movements.
3. Fingerspelling recognition, which focuses on identifying sequences of manually spelled letters, typically using a fixed set of characters from the language alphabet.

Different input modalities can be used for SLR, such as RGB video for capturing appearance, depth maps for 3D spatial information, and skeleton data for tracking body and hand keypoints (Sarhan et al., 2023).

Early approaches, given the sequential nature of the problem, focused on modeling handshape and hand movement features using Hidden Markov Models (HMMs) (Vogler et al., 1998; Vogler et al., 1999; Cooper et al., 2011), which are well-suited for capturing temporal dynamics in gesture sequences. With advances in deep learning, methods have shifted towards Convolutional Neural Network (CNN)-based architectures (Camgoz et al., 2017; Adaloglou et al., 2021), where 2D CNNs extract spatial features from individual video frames and Long Short-Term Memory (LSTM) networks capture temporal dependencies across the signing sequence. Alternatively, 3D CNN-based approaches process spatio-temporal features directly from video data (Sarhan et al., 2020; Mercanoglu et al., 2022; Papadimitriou and Potamianos, 2023a), capturing motion and appearance jointly. Building on the success of transformer models in Natural Language Processing, transformer-based architectures have also been adapted for SLR, leveraging self-attention mechanisms to model long-range dependencies in signing sequences (Camgoz et al., 2020; De Coster et al., 2020).

### 2.1.2 Sign language translation

While SLR involves mapping visual input to glosses (a written representation of sign language units), this output remains within the domain of sign language. In contrast, SLT extends this task by translating the recognized glosses (or visual input) into spoken language text or speech, thereby bridging the gap between sign and spoken modalities. This task is particularly challenging due to the differences in grammar, syntax, and semantics between sign languages and spoken languages. Early approaches often used glosses (either predicted or manually annotated) as an intermediate representation before translating them into spoken language text (Moryossef et al., 2021; Müller et al., 2023). More recent end-to-end approaches leverage Neural Machine Translation techniques to convert sequences of visual features directly into spoken language text (Camgoz et al., 2018; Camgoz et al., 2020; Zhou et al., 2023). These

methods typically adopt an encoder-decoder architecture, where the encoder processes the visual input and the decoder generates the corresponding spoken language tokens.

SLR and SLT are often used as automatic sign language interpretation technologies, bridging the communication gap between hard-of-hearing and hearing individuals. SLT systems can also be used for real-time video subtitling, where the sign language input is translated into spoken language text for live events or media content (Bull et al., 2021).

### 2.1.3 Sign language generation

SLG focuses on generating sign language content from various input representations. These inputs can include spoken language text, sign language glosses, HamNoSys annotations, or even other sign videos in the context of data augmentation (Stoll et al., 2018; Stoll et al., 2020; Arkushin et al., 2023; Moryossef, 2024). The output of such systems may take the form of 3D avatars (Neves et al., 2020), photorealistic RGB videos, or sequences of 2D/3D pose keypoints (Baltatzis et al., 2024), depending on the target application and the rendering technique employed.

SLG plays a critical role in improving accessibility and enabling communication for Deaf and hard-of-hearing individuals. SLG systems can function as automatic interpreters, converting spoken or written language content into sign language. Moreover, the generated sign sequences can serve as reference material for sign language learners, providing model examples for improving sign production and comprehension.

### 2.1.4 Sign language assessment

SLA aims to evaluate the quality and correctness of sign language production with respect to a reference sign, often in the context of sign language learning. Unlike recognition, which focuses on identifying the sign, assessment systems also aim to determine how well a sign is produced in terms of linguistic features such as handshape accuracy, movement clarity, location precision, facial expressions, etc.

One of the ways to bridge the communication gap between the Deaf and hard-of-hearing community and the Hearing community is to develop assistive technology that can help people, irrespective of whether they are hearing impaired to learn sign language, assess and improve themselves with the help of automatic systems that provide meaningful feedback. In that direction, there has been effort for more than a decade in developing interactive sign language learning platforms for both children and adults (Spaai et al., 2005; Brashear et al., 2006; Aran et al., 2009; Zafrulla et al., 2011).

Most of the existing platforms for sign language learning or assessment test vocabulary using pre-recorded videos for later analysis. E-learning platforms such as SignAssess (John, 2012) allow to compare a user's recorded video to a reference one. In terms of real-time sign language

verification, *SignAll* (SignAll Technologies Inc., 2021) and *ISARA* (ISARA, 2016) applications assess if the produced sign is correct or incorrect. Validation in terms of whether the produced sign is correct or incorrect is not enough information to aid a sign language learner. The SL-ReDu (Papadimitriou, Potamianos, et al., 2023b) platform introduces a signer-independent Greek SLR system for learner assessment, integrating deep learning-based visual processing to handle both isolated signs and finger spelled sequences in real-world webcam videos. On the linguistics side, Willoughby et al. (2015) envisioned a system, *My Interactive Auslan Coach*, that provides automatic feedback on correctness of handshape and hand movement of Australian Sign Language signs. Huenerfauth et al. (2017) proposed a system that analyzes the production of signs and gives feedback with respect to both manual and non-manual components. These two systems just depict the prototype of the feedback system and analyze the system in terms of usability through a wizard-of-oz setup.

Recent works have focused on developing automated SLA systems that provide interpretable feedback on sign production (Tornay, Camgoz, et al., 2020; Cory et al., 2024). These systems typically use machine learning techniques to analyze the signing performance and identify errors in specific channels, such as handshape, movement trajectory, etc. The goal is to provide learners with detailed feedback that aligns with how human instructors would assess the signing performance.

SLA plays a crucial role in sign language education by providing actionable feedback that helps learners improve their signing skills. It is also essential for the development of SLG systems, offering a means to evaluate the linguistic and articulatory quality of the generated signs, as we see later in Chapter 5.

## 2.2 Methods

This section provides an overview of the methods that underpin the experimental studies in this thesis. These components serve as the basis for our proposed evaluation frameworks, comparisons, and system designs across different SLA tasks.

In the thesis by Tornay (2021), it has been shown that signs can be assessed over different channels of production like handshape, hand movement, etc. Two levels of assessment have been proposed:

1. **Lexeme-Level**: Whether the produced sign matches the reference sign.
2. **Form-Level**: Feedback on whether the form of the sign in terms of handshape, hand movement, etc. is correct or incorrect.

Assessment that goes beyond simply determining whether a sign is correct or incorrect is crucial both for sign language learning and for evaluating generated content, since even minor variations in articulatory channels can affect communication and alter the intended

interpretation of a sign.

The remainder of the section is organized as follows: In Section 2.2.1, we introduce the phonology-based approach used to develop reference models for SLA. In Section 2.2.2, we present the assessment framework based on these models.

### 2.2.1 Phonology-based sign language recognition

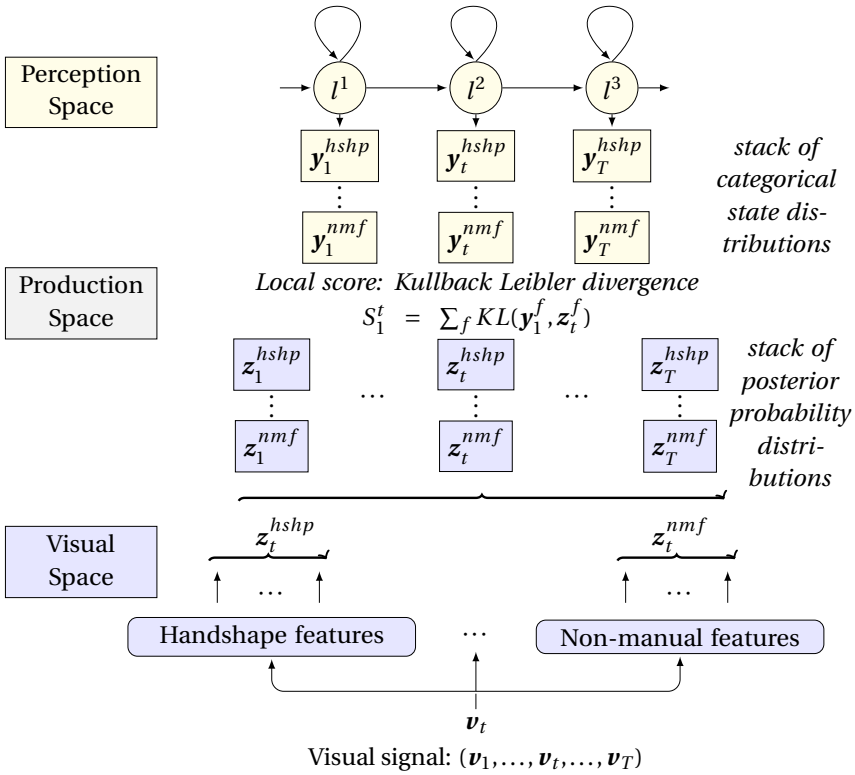


**Figure 2.2** – Illustration of modeling production and perception phenomena in KL-HMM framework for sign language processing. The visual signal is denoted by $(v_1, v_2, \ldots, v_T)$, $\left[\boldsymbol{z}_t^{hshp}, \ldots, \boldsymbol{z}_t^{hmvt}, \ldots, \boldsymbol{z}_t^{nmf}\right]$ is the stack of posterior estimates of channels obtained from the visual signal at time $t$, and the emission distribution for HMM state $n$ is parameterized by the categorical distribution $\left[\boldsymbol{y}_n^{hshp}, \ldots, \boldsymbol{y}_n^{hmvt}, \ldots, \boldsymbol{y}_n^{nmf}\right]$.

Sign Language like spoken language has a production and perception component.

- *Production* involves the generation of the signal. In speech, it is the movement of articulators such as vocal folds, tongue, jaw, and lips that produce a one-dimensional acoustic signal. In contrast, in sign language, a two-dimensional visual signal is generated through articulations involving hand movements, facial expressions, and body postures.
- *Perception* involves interpreting the signal using linguistic units such as words or phrases.

> In spoken language, the signal is interpreted through well-defined auditory subunits such as phonemes, syllables and words. In sign language, the linguistic units are not as well-defined and the concept of subunits as a combination of different articulators is still being explored.

Owing to the similarities between spoken and sign language, the posterior-based articulatory feature modeling for speech processing (Rasipuram et al., 2016) was extended to SLP (Tornay, Razavi, et al., 2019).

Posterior-based approaches model articulatory features as categorical distributions and use Kullback-Leibler Hidden Markov Models (KL-HMMs) (Aradilla et al., 2007; Aradilla et al., 2008) to capture their temporal dynamics. In this probabilistic framework, a sequence of posterior distributions over subunits is first computed from the input visual signal $\boldsymbol{v}$ of length $T$. At each time step $t$, the posterior vector is given by:

$$\boldsymbol{z}_t = \left[\boldsymbol{z}_t^{hshp}, \boldsymbol{z}_t^{hmvt}, \boldsymbol{z}_t^{nmf}\right]$$

where $\boldsymbol{z}_t^{hshp}, \boldsymbol{z}_t^{hmvt}$, and $\boldsymbol{z}_t^{nmf}$ correspond to the posterior distribution of handshape, hand movement and non-manuals respectively.

These stacks of posterior distributions are then used to train an HMM whose states $n \in [1, N]$ are each parameterized by a corresponding categorical distribution:

$$\boldsymbol{y}_n = \left[\boldsymbol{y}_n^{hshp}, \boldsymbol{y}_n^{hmvt}, \boldsymbol{y}_n^{nmf}\right]$$

The three possible cases to define the local scores based on KL-divergence between $\boldsymbol{y}_i$ and $\boldsymbol{z}_t$ are given by:

1. KL-divergence (KL):

$$S_{KL}(\boldsymbol{y}_i, \boldsymbol{z}_t) = \sum_f \sum_{d=1}^{D_f} \boldsymbol{y}_t^{f,d} \log\left(\frac{\boldsymbol{y}_t^{f,d}}{\boldsymbol{z}_i^{f,d}}\right) \tag{2.1}$$

2. Reverse KL-divergence (RKL):

$$S_{RKL}(\boldsymbol{y}_i, \boldsymbol{z}_t) = \sum_f \sum_{d=1}^{D_f} \boldsymbol{z}_t^{f,d} \log\left(\frac{\boldsymbol{z}_t^{f,d}}{\boldsymbol{y}_i^{f,d}}\right) \tag{2.2}$$

3. Symmetric KL-divergence (SKL):

$$S_{SKL}(\boldsymbol{y}_i^f, \boldsymbol{z}_t^f) = \frac{1}{2}(S_{KL}(\boldsymbol{y}_i, \boldsymbol{z}_t) + S_{RKL}(\boldsymbol{y}_i, \boldsymbol{z}_t)) \tag{2.3}$$

where $D_f$ corresponds to the dimension of the channel $f$ and $\boldsymbol{z}_t^{f,d}$, $\boldsymbol{y}_i^{f,d}$ are the $d$-th elements of the posterior and state distributions respectively.

The HMM parameters are estimated using the Viterbi Expectation-Maximisation (EM) algorithm by minimizing the score based on KL-divergence between the input posterior sequence and the state distributions, aligning the predicted articulatory behavior with the observed signal. The decoding is performed using standard Viterbi decoder using the KL-divergence based local score. In our experiments, we use the score based on the Reverse KL-divergence (Equation 2.2).

The overall framework is illustrated in Figure 2.2. Posterior features for individual channels can be derived using various approaches, including both handcrafted feature extraction and deep learning-based methods. The framework is inherently flexible and can be extended to incorporate any number of channels. The work presented in (Tornay, 2021) specifically focuses on **manual** features extracted from 3D skeletons obtained via Kinect sensors.

### 2.2.2 Phonology-based sign language assessment

In the context of SLA, KL-HMMs can be interpreted as template models that represent canonical or reference sign productions. Assessment is then formulated as a sign matching problem, where a test sign production is compared to the reference model using Dynamic Time Warping (DTW) via the Viterbi algorithm. This comparison determines whether the produced sign is acceptable based on its alignment with the reference. The local constraints for DTW are defined analogously to the left-to-right transition structure of HMMs, enforcing a temporal progression through the sign. Given the reference model, the posterior feature sequence of a test production is aligned to it using DTW, and the global and local scores are thresholded to determine the acceptability of the produced sign (Tornay, Camgoz, et al., 2020).

The assessment approach is illustrated in Figure 2.3. The method matches the stack of posterior sequences from the produced sign video $\boldsymbol{z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T]$ with the sequence of KL-HMM states of the expected sign characterized by categorical distributions $\boldsymbol{y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N]$, where $n$ is the state. The local score given by $\mathrm{l}(\boldsymbol{y}_n, \boldsymbol{z}_t)$ is based on Symmetric KL-divergence (Equation 2.3) of two probability distributions. The match is obtained by dynamic programming with the following recursion:

$$\mathrm{S}(n,t) = \mathrm{l}(\boldsymbol{y}_n, \boldsymbol{z}_t) + \min\left[\mathrm{S}(n,t-1) + c_x, \mathrm{S}(n-1,t-1) + c_x\right]$$

where $c_x = -\log(0.5)$ is the transition cost and $\mathrm{l}(\boldsymbol{y}_n, \boldsymbol{z}_t)$ is the local score given by:

$$\mathrm{l}(\boldsymbol{y}_n, \boldsymbol{z}_t) = \sum_f \mathrm{SKL}\left(\boldsymbol{y}_n^f, \boldsymbol{z}_t^f\right)$$

$$\mathrm{SKL}(\boldsymbol{y}_n^f, \boldsymbol{z}_t^f) = \frac{1}{2}\sum_{d=1}^{D_f} \boldsymbol{y}_n^{f,d} \log\left(\frac{\boldsymbol{y}_n^{f,d}}{\boldsymbol{z}_t^{f,d}}\right) + \boldsymbol{z}_t^{f,d} \log\left(\frac{\boldsymbol{z}_t^{f,d}}{\boldsymbol{y}_n^{f,d}}\right) \tag{2.4}$$

where $D_f$ corresponds to the dimension of the channel $f$ and $\boldsymbol{z}_t^{f,d}$, $\boldsymbol{y}_n^{f,d}$ are the $d$-th elements

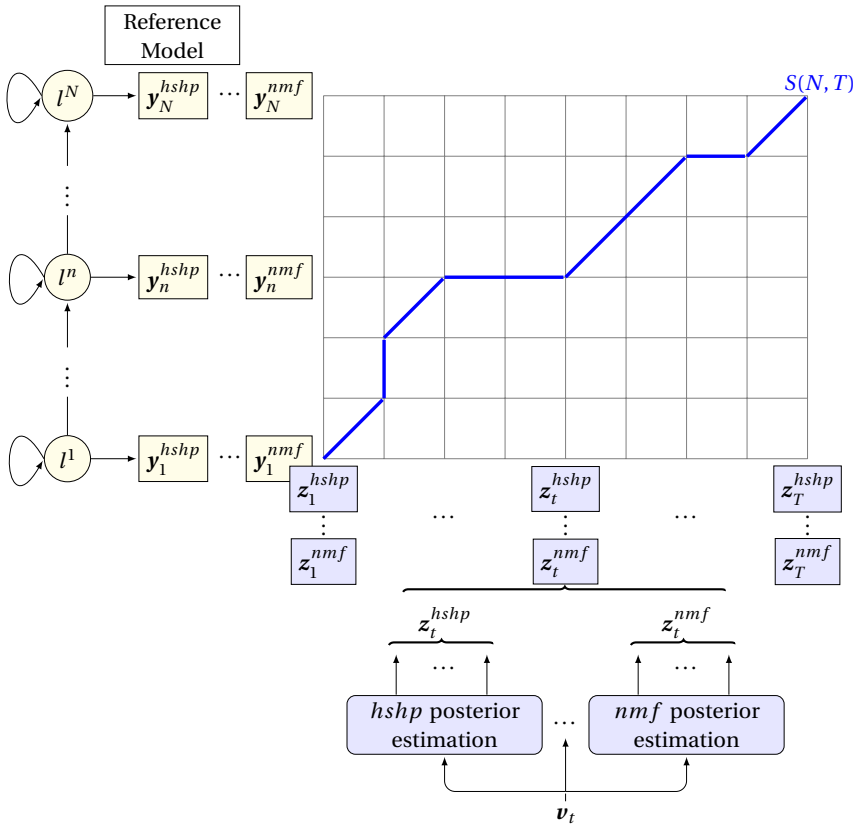


**Figure 2.3** – Illustration of the assessment framework. $\left[z_t^{hshp} \ldots z_t^{nmf}\right]$ is the stack of posterior estimates of the visual sub-units obtained from the test signer production at time $t$. Each state $l_n$ of the reference KL-HMM model is parameterized by the categorical distribution $\left[y_n^{hshp} \ldots y_n^{nmf}\right]$. The DTW score is given by $S(N,T)$.

of the posterior and state distributions, respectively.

Given an input sequence of posterior features, we compute the alignment to the reference using DTW based on the cost function described above. This results in the best matching path that maps each HMM state $n \in [1,\ldots,N]$ to a segment in time defined by its beginning and end frame indices $\left(t_n^b, t_n^e\right)$. These aligned segments are then used to compute both lexeme-level and form-level assessment scores.

Given the best matching path $(t_n^b, t_n^e)$ for each state $n$, the score for lexeme-level assessment is calculated as:

$$S_{lex} = \frac{1}{N} \sum_{n=1}^{N} \frac{\sum_{t_n^b}^{t_n^e} l(y_n, z_t)}{t_n^e - t_n^b + 1}$$

As we can separate out the contributions of each of the channels from the global score, the

form-level assessment scores for each channel can be factored from this as;

$$S^{f}_{form} = \frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{t_n^b}^{t_n^e}\mathrm{SKL}(y_n^f, z_t^f)}{t_n^e - t_n^b + 1}$$

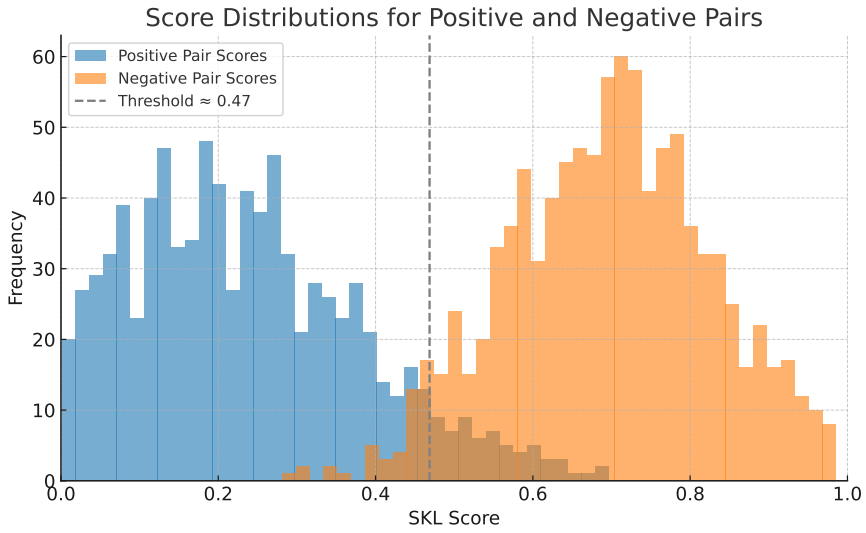


**Figure 2.4** – Histogram of SKL scores of positive and negative sign pairs. The line shows the decision boundary for the acceptability decision.

To determine whether a produced sign instance is acceptable, we apply a threshold to the computed assessment scores $S_{form}$ and $S_{lex}$. This threshold is learned by evaluating the system on a development set consisting of positive pairs (instances of the same sign class) and negative pairs (instances of different sign classes). Assessment scores are computed for each pair using the lexeme-level or form-level scoring formulation. A threshold is then selected to maximize the F1 score for the binary acceptability decision, effectively balancing precision and recall in distinguishing correct from incorrect productions. Figure 2.4 illustrates the decision boundary selection.

These methods have been validated on SLA using 3D skeleton data from a Kinect-v2 sensor (Tornay, Camgoz, et al., 2020). However, reliance on such specialized hardware limits scalability and accessibility, particularly in real-world learning scenarios where affordable and widely available sensors (e.g., webcams) are preferred. To address this limitation, Chapter 3 and Chapter 4 explore the use of RGB-based pose estimation methods for assessment.

## 2.3 Datasets

In this section, we provide an overview of the datasets used in our experiments. These include datasets for isolated sign language assessment, alignment of continuous signing sequences, and detection of non-manual features.

### 2.3.1 SMILE-DSGS

The **SMILE-DSGS** (Deutschschweizerische Gebärdensprache, DSGS) dataset (Ebling et al., 2018) was created to develop an assessment system for the lexicon of Swiss German Sign Language. It is the *only* database that has *linguistically annotated* sign language data to aid production-level SLA. The dataset is composed of 100 isolated signs from DSGS. The data was acquired from 11 adult L1 (first language) signers and 19 adult L2 (second language) learners performing the signs of a DSGS vocabulary production test. The inclusion of both L1 and L2 signers significantly enhances the diversity of the dataset, capturing variations in signing styles influenced by linguistic proficiency and individual differences. The videos were collected with a Microsoft Kinect v2 sensor, and the dataset includes both RGB and depth data obtained by the sensor and the gloss (meaning label associated with the sign in related spoken language) annotations. The linguistic annotations capture the variations and the acceptability of the signs through six categories, based on linguistic criteria (lexeme, meaning, and form). The category evaluates the acceptance of the produced sign according to whether it is the same lexeme (word), has the same meaning, and has the same form as the target sign. This comprehensive annotation approach makes the SMILE-DSGS dataset a valuable resource for capturing the nuanced variability in sign production and supporting advanced SLA. The categories are defined as follows:

1. Category 1 - Same lexeme as target sign: same meaning, same form;
2. Category 2 - Same lexeme as target sign: same meaning, slightly different form;
3. Category 3 - Same lexeme as target sign: same meaning, different form;
4. Category 4 - Same lexeme as target sign: slightly different meaning, slightly different form;
5. Category 5 - Different lexeme than target sign: same meaning, different form;
6. Category 6 - Different lexeme than target sign: different meaning, different form;

**Table 2.1** – Acceptability levels and number of samples for different signer categories.

| Category | Acceptability | #Samples |
|---|---|---|
| Cat 1, Cat 2 | High | 581 |
| Cat 3, Cat 4 | Medium | 412 |
| Cat 5, Cat 6 | No | 183 |

The acceptable productions of Category 1 and 2 were partitioned in a signer-independent manner into 987 training samples from 15 signers, 509 validation samples from 7 signers, and 581 test samples from 8 signers. These acceptable productions are used to build the reference models for assessment. The data from the other categories (3-6) with statistics as shown in Table 2.1 are used to evaluate the assessment system.

### 2.3.2 Aff-Wild2

Aff-Wild2 (Zafeiriou et al., 2017; Kollias, Tzirakis, et al., 2019; Kollias, Sharmanska, et al., 2019; Kollias and Zafeiriou, 2019; Kollias and Zafeiriou, 2021a; Kollias, Sharmanska, et al., 2021; Kollias et al., 2020; Kollias and Zafeiriou, 2021b; Kollias, 2022; Kollias, 2023a; Kollias et al., 2023b) is a large-scale audiovisual dataset developed for affective behavior analysis in unconstrained, real-world conditions. Released as part of the Affective Behavior Analysis in the Wild (ABAW) competition series, the dataset comprises 564 videos totaling approximately 2.6 million frames, capturing subjects expressing spontaneous emotions across a wide range of settings. It features extensive diversity in age, gender, ethnicity, nationality, and environment, making it a robust benchmark for real-world affective computing tasks. Out of the 564 videos in the Aff-Wild2 dataset, 295 videos are used for training, 105 videos for validation, and 164 videos are held out for testing (no labels released), following the official data split protocol.

Aff-Wild2 provides frame-level annotations for:

- valence and arousal (continuous emotion dimensions),
- the seven basic expressions: happiness, surprise, anger, disgust, fear, sadness, and neutrality, and
- facial action units (FAUs) defined according to the Facial Action Coding System (FACS).

**Table 2.2** – FAUs and their associated face regions annotated in Aff-Wild2 dataset.

| # | AU Code | Description | Face Region |
|---|---|---|---|
| 1 | AU1 | Inner Brow Raiser | Brows |
| 2 | AU2 | Outer Brow Raiser | Brows |
| 3 | AU4 | Brow Lowerer | Brows |
| 4 | AU6 | Cheek Raiser | Cheeks |
| 5 | AU7 | Lid Tightener | Eyes |
| 6 | AU10 | Upper Lip Raiser | Mouth |
| 7 | AU12 | Lip Corner Puller | Mouth |
| 8 | AU15 | Lip Corner Depressor | Mouth |
| 9 | AU23 | Lip Tightener | Mouth |
| 10 | AU24 | Lip Pressor | Mouth |
| 11 | AU25 | Lips Part | Mouth |
| 12 | AU26 | Jaw Drop | Jaw |

Due to its dense frame-level annotations over entire video sequences, Aff-Wild2 is well-suited for both frame-wise and spatio-temporal modeling approaches. This makes it particularly valuable for tasks such as FAU detection in dynamic, real-world conditions. In this thesis, we leverage the Aff-Wild2 dataset to train facial action unit detection models, which are then transferred to the domain of non-manual feature detection in sign language videos. Table 2.2

lists the FAUs annotated in the Aff-Wild2 dataset, which are used to train the FAU detection models.

### 2.3.3 SMILE-SRT

The SMILE Sentence Repetition Test (SRT) dataset (Cory et al., 2024) was created to support the development of assessment systems for continuous Swiss German Sign Language (DSGS) productions. During data collection, each participant viewed a prerecorded video of a DSGS sentence twice and was then asked to reproduce what they had seen. The dataset was constructed by recording a repetition test across 12 sentences of varying difficulty, determined by the number of signs as well as their morphological and syntactic complexity.

This dataset was developed as part of the SMILE-II project, which aims to extend sign language learning and assessment technologies from isolated signs to continuous sign sequences. A major challenge in this transition is the lack of continuous sign language datasets with rich linguistic annotations spanning phonological, morphological, lexical, semantic, syntactic, and prosodic levels. The SMILE-SRT dataset specifically addresses this gap by providing a linguistically annotated dataset suitable for phonologically grounded assessment in continuous signing scenarios.

The collected dataset comprises 104 productions signed by nine deaf native signers/early learners (L1 signers) and 100 productions signed by 14 learners (L2 signers) enrolled in an interpreter training program. Of the nine L1 signers, six identified as female and three as male (mean age 33.9; range 20 – 49), while all the L2 signers identified as female (mean age 34.8; range = 22 – 46). The dataset was gathered within a study environment equipped with 5 cameras (three machine vision cameras and two webcams) from different angles. In this work, we use data only from the front-facing camera. Participants were shown a pre-recorded signed sentence twice in a row with the task of signing it in the same way, copying both manual and non-manual activities. The videos were recorded at 57 frames per second (fps) and a resolution of 1024x1024 pixels in RGB.

**Table 2.3** – Non-manual categories annotated in the SMILE SRT dataset.

| Region | Subcategory | States |
|---|---|---|
| Face | Cheeks | – |
| | Eyebrows | Firm, Relaxed |
| | Eyelids | Closed, Open, Partial |
| | Gaze | Addressee, Downward+Level, OnHands, Upward |
| | Lips | Closed+Firm, Closed+Relaxed, Slightly Open+Relaxed |
| | Nose | Static |
| Head | Position | Back, Front, Lateral |
| | Movement | Nod, Shake, StrongDynamic |
| Upper Body | Shoulders | Raised |
| | Torso | Upright, Oriented |

The dataset is comprehensively annotated with both manual and non-manual features relevant to sign language production. Manual annotations include segmentation of individual signs, identified by their start and end timestamps within continuous sentence-level signing. This segmentation allows for isolated analysis and evaluation of each sign instance. In parallel, non-manual features listed in Table 2.3 are also annotated with precise temporal boundaries, indicating when each non-manual signal begins and ends. Notably, there is considerable variation between the non-manual features annotated in this dataset (Table 2.3) and those present in the FAU dataset (Table 2.2).

In addition to manual and non-manual annotations, the dataset includes human ratings provided by trained evaluators. Raters are instructed using a standardized rubric and assess videos at the sentence level across six linguistic criteria: manual components, mouth movements, eyebrow movements, head movements, eye gaze, and overall sentence structure. Each criterion is evaluated on a three-point scale, allowing for more nuanced and graded feedback compared to binary assessments. This enriched annotation framework supports the assessment of both manual and non-manual features, enabling a comprehensive assessment of continuous sign language production.

We use this dataset for studies on alignment of continuous signing sequences and for non-manual feature detection.

# 3 Accessible sign language assessment using webcam based systems

**Chapter Overview:** This chapter investigates the feasibility of webcam-based sign language assessment by replacing depth sensor input with RGB-based 2D or 3D skeletons. We evaluate how this shift impacts assessment and recognition performance within the phonology-based framework.

This work is situated within the broader goal of developing sign language production assessment systems to support sign language learning. Sign language learning requires not only recognizing signs but also providing constructive feedback on how signs are produced. Recent work on phonology-based SLA (Tornay, Camgoz, et al., 2020) demonstrated that modeling signs as sequences of handshape and hand movement subunits within the phonology-based framework enables both lexeme-level and form-level (handshape, movement, etc.) assessments that are interpretable and linguistically grounded. This framework was validated using 3D skeleton data obtained from Kinect sensors. However, the reliance on such specialized hardware limits the accessibility and scalability of the system, particularly in at-home or resource-constrained learning environments.

To address these limitations, we explore the feasibility of building accessible SLA systems that operate on RGB video input. Webcams are widely available and embedded in most personal devices, making them a practical alternative to depth sensors. However, transitioning from 3D skeleton data to skeletons estimated from RGB videos introduces several challenges: the loss of true depth information, variability in skeleton estimation quality, and potential degradation in recognition and assessment performance.

In this chapter, we investigate these challenges through the following research questions (RQ):

- **RQ 1**: *What is the impact of loss of depth information on sign language production assessment?*

  To address this, we evaluate the performance of various RGB-based skeleton estimation approaches—including OpenPose (Cao et al., 2019) and Mask R-CNN (He et al., 2017)

(2D)—and compare them against a Kinect-based baseline in terms of lexeme and form-level assessment scores. This quantifies the impact in assessment quality due to the absence of true depth data.

- **RQ 2**: *Can the impact of loss of depth information on sign language production assessment be mitigated using 3D joint estimation techniques applied to RGB video?*

  To address this, we compare 2D methods with RGB-based 3D skeleton estimation techniques such as MediaPipe (Lugaresi et al., 2019), VideoPose3D (Pavllo et al., 2019), and VIBE (Kocabas et al., 2020) to determine whether 3D joint estimation can effectively compensate for the lack of true depth and restore assessment performance.

- **RQ 3**: *Given that SLA inherently involves SLR components, what is the trade-off between assessment performance and recognition performance when using RGB-derived skeletons?*

  To address this, we conduct an SLR study using the same reference models trained for assessment and skeleton features from each method, analyzing how changes in recognition performance relate to changes in assessment accuracy when depth is unavailable.

We conduct all the studies using the SMILE-DSGS dataset—the only linguistically annotated dataset available for sign language production assessment in Swiss German Sign Language.

The remainder of this chapter is organized as follows: Section 3.1 presents an overview of the RGB-based pose estimation methods used in this study, including both 2D and 3D approaches. Section 3.2 describes the experimental setup for posterior feature extraction for reference model training and model configurations. In Section 3.3, we report and analyze the results of SLR and form-level assessment across different pose estimation techniques. Finally, Section 3.4 summarizes the contributions of this chapter.

All results presented in this chapter except those involving MediaPipe have been previously published in (Tarigopula et al., 2022).

## 3.1 Skeleton estimation approaches

We explore a range of skeleton estimation techniques that either extract 2D or 3D skeletons from RGB images or videos. These methods differ in their architectures, the way temporal information is used and inference strategy. We categorize them into two main groups: 2D skeleton estimation methods and 3D skeleton estimation methods, and we provide a brief technical overview of each.

### 3.1.1 2D skeleton estimation

#### OpenPose

OpenPose (Cao et al., 2019) is a widely used framework for real-time multi-person 2D pose estimation. It follows a bottom-up approach, detecting all body parts in the image first and then grouping them into individual person instances. The method introduces Part Affinity Fields (PAFs), which are 2D vector fields that encode both the position and orientation of limbs by capturing the spatial relationships between keypoints. A CNN-based architecture is used to jointly predict confidence maps for body parts and the corresponding PAFs, which are then parsed to assemble full-body skeletons for each detected individual.

#### Mask R-CNN

Mask R-CNN (He et al., 2017) is a widely used instance segmentation framework that can be extended to 2D pose estimation. It employs a convolutional backbone to extract image features and follows a two-stage architecture: first, it generates region proposals using a Region Proposal Network, and then it classifies and refines the features obtained from the proposals to produce masks. For 2D keypoints estimation, it treats each keypoint as a one-hot encoded mask and predicts one mask for each keypoint.

### 3.1.2 3D skeleton estimation

#### MediaPipe

MediaPipe employs a two-stage approach to pose estimation. It first detects the presence and location of a person in the image, and then predicts the 3D coordinates of body landmarks using the BlazePose model (Bazarevsky et al., 2020). To reduce inference cost, MediaPipe avoids standard heatmap-based decoding; instead, it directly regresses 3D keypoint coordinates, using heatmap and offset supervision during training. The heatmap branch is discarded at inference time, enabling efficient real-time performance on resource-constrained devices.

#### VideoPose3D

VideoPose3D (Pavllo et al., 2019) uplifts 2D keypoints to 3D by leveraging dilated temporal convolutions. It takes as input a sequence of 2D keypoints and predicts 3D joint locations frame-by-frame. The model learns temporal patterns by applying 1D convolutions to enable smooth and plausible 3D poses. The model is trained on a large corpus of 3D poses from the Human 3.6M dataset (Ionescu et al., 2014).

#### Video inference for body pose and shape estimation (VIBE)

VIBE (Kocabas et al., 2020) is a method for 3D human pose and shape estimation from videos. It predicts 3D human pose and shape by regressing Skinned Multi-Person Linear Model (SMPL) parameters from frame sequences. It uses a CNN backbone to extract frame-level features, followed by a temporal encoder that outputs SMPL parameters. Owing to lack of 3D annotated motion data, VIBE leverages adversarial learning to discriminate real human motions from the ones generated by temporal shape and pose regressors to predict reliable 3D SMPL (Loper et al., 2015) human models from in-the-wild videos.

## 3.2 Experimental setup

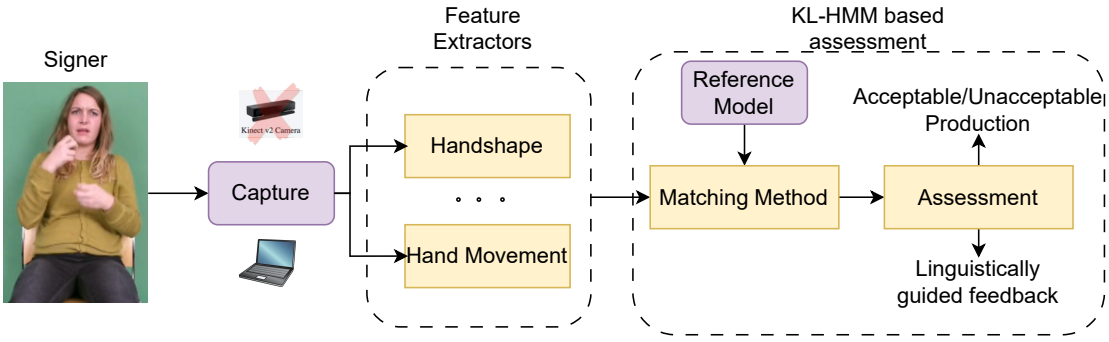


**Figure 3.1** – Overview of the accessible sign language assessment framework.

In Figure 3.1, we present the schematic overview of the SLA framework. The framework consists of two main components: (i) Feature Extraction and (ii) KL-HMM-based SLA. The feature extraction component is responsible for estimating the hand movement and handshape posteriors from the input video. The KL-HMM-based SLA component uses these posteriors to perform lexeme-level and form-level assessments.

### 3.2.1 Feature extraction

#### Hand movement posterior extraction

To estimate hand movement posterior features $z_t^{hmvt}$, we adopt the subunit extraction approach proposed in (Tornay and Magimai.-Doss, 2019), which models the dynamic structure of sign movements as a sequence of underlying motion subunits. The posterior probability distribution over these subunits is used as the hand movement posterior at each frame. The process involves two main steps:

1. **Hand movement subunit inference**: For each sign instance, a sequence of hand movement feature vectors is extracted based on skeletal data. Each feature vector encodes the coordinates of the left and right hands relative to key reference joints (head, hip, and

shoulder), along with their velocities. To account for inter-signer variability, the skeletal data is normalized by aligning neck joints to a reference signer and scaling by shoulder width. Given the extracted sequences, hand movement subunits are inferred by training left-to-right HMM with Gaussian Mixture Models (GMM) emissions. For each sign, several HMMs with varying numbers of states are trained, the optimal number of states is selected based on recognition performance on a development set, with each HMM state representing a distinct motion unit within the sign.

2. **Posterior probability estimation**: Once the subunit alignments are obtained between the input sequences and the states, a multilayer perceptron (MLP) is trained to classify each frame into its corresponding motion subunit. The MLP takes as input a context window of nine consecutive frames of skeletal features and is trained using a cross-entropy loss. Its architecture is selected through cross-validation. The resulting posterior features $\boldsymbol{z}_t^{hmvt}$ represents the likelihood of each motion subunit over time, enabling structured and interpretable modeling of hand movement dynamics.

Figure 3.2 shows the illustration of the subunit based posterior extraction.

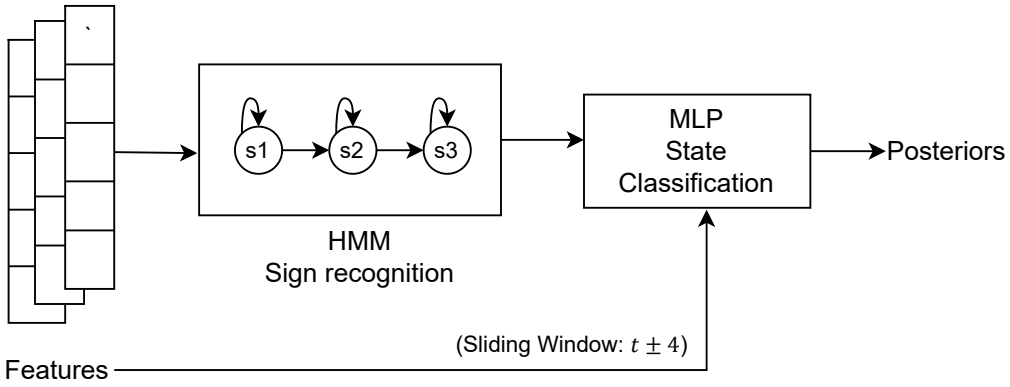


**Figure 3.2** – Subunit-based hand movement posterior extraction.

### Handshape posterior extraction

To estimate the frame-level handshape posterior features $\boldsymbol{z}_t^{hshp}$, we employ the SubUNets approach proposed in (Camgoz et al., 2017), which is based on a CNN-BiLSTM sequence-to-sequence architecture. The model takes as input a sequence of cropped hand patches and outputs a sequence of handshape class probabilities. Spatial features are first extracted from each frame using a CNN backbone. These features are then processed by a Bidirectional Long Short-Term Memory (BiLSTM) network to capture temporal dynamics across the sequence. The model is trained using a Connectionist Temporal Classification (CTC) loss, which enables it to simultaneously recognize and align handshape sequences without requiring explicit frame-level annotations.

For posterior estimation, we use a pretrained SubUNets model trained on the One-Million Hands dataset (Koller et al., 2016), which covers 61 handshape classes, including a blank label. Given a video of hand patches, the model outputs frame-wise posterior distributions over these handshape classes, which we use as the handshape posterior features $\boldsymbol{z}_t^{hshp}$.

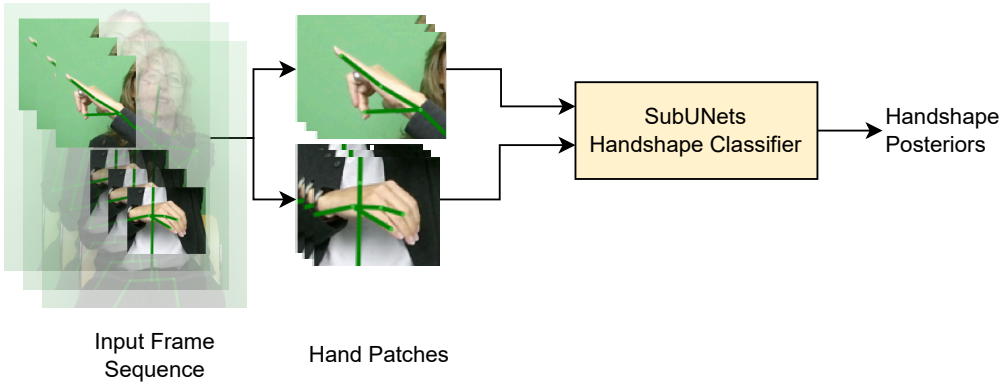


**Figure 3.3** – Handshape posterior extraction.

### 3.2.2 KL-HMM systems

Following feature extraction, we configure multiple KL-HMM systems to model isolated signs using the extracted handshape and hand movement posterior features. Each KL-HMM models a single sign class using frame-level posterior features as observations, with emission probabilities represented as categorical distributions over subunit posteriors as described in Section 2.2.1. These models serve as reference models in the assessment framework, against which test sign productions are evaluated. We investigate several system variants based on how the input features are constructed:

1. **M**: This system models only the hand movement channel. Hand movement subunits are inferred using skeletal features from the combined motion of both hands, and the resulting posteriors are used to train the reference model.

2. **rlM**: In contrast to **M**, this system models hand movement separately for the right and left hands. Subunits are inferred independently for each hand, and two separate MLPs are trained to estimate the corresponding posteriors. These posteriors are then concatenated and used to train the reference model.

3. **M+rlS**: This system combines the hand movement posteriors from the **M** configuration (i.e., combined-hand movement subunits) with handshape posteriors. The two channels are concatenated and jointly modeled.

4. **rlM+rlS** This is the most detailed system, in which hand movement posteriors for the right and left hands (as in **rlM**) are concatenated with handshape posteriors, allowing finer-grained modeling across both manual channels.

The reference models are trained exclusively on linguistically acceptable sign productions, corresponding to Category 1 and 2 of the SMILE-DSGS dataset. These categories represent correct or near-correct realizations of a target sign in terms of both meaning and form. For each sign class, KL-HMMs with varying numbers of states (ranging from 3 to 30) are trained, and the model yielding the best isolated sign recognition accuracy on the development set is selected. For SLA, we follow the procedure described in Section 2.2.2 to determine thresholds for lexeme-level and form-level evaluations (hand movement and handshape). These thresholds are derived from the development set by computing KL divergence-based similarity scores between pairs of acceptable sign instances (for correct matches) and between different signs (for incorrect matches). The thresholds are then chosen to maximize the $F1$ score on the development data for each type of assessment.

Besides SLA studies, to address RQ 3, we carried out SLR studies.

## 3.3 Results and analysis

In this section, we present and analyze the experiment results that were performed to address the research questions presented earlier.

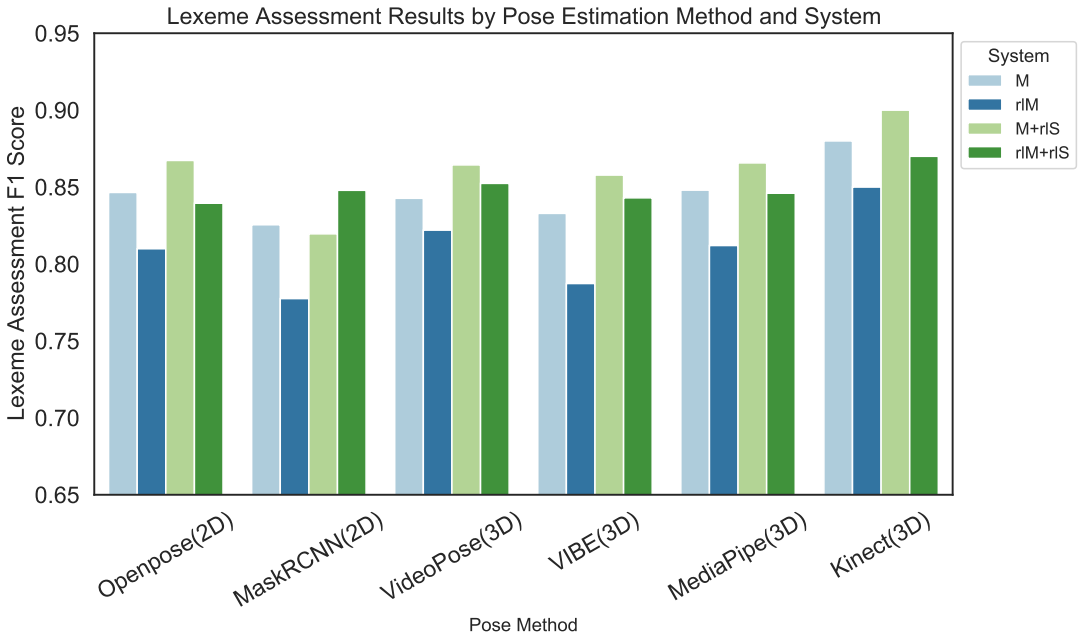


**Figure 3.4** – Lexeme assessment F1 scores.

Figure 3.4 shows the lexeme-assessment F1 scores across different skeleton estimation methods and KL-HMM configurations. The Kinect, using true depth information, consistently achieves the best performance, serving as a strong baseline. Among the RGB-based 3D skeleton estimation methods, VideoPose3D and MediaPipe yield very similar and competitive

performance levels, closely approaching the Kinect-based system. In contrast, VIBE shows significantly lower performance, highlighting the challenges it faces in accurately estimating skeletal data suitable for SLA. Among 2D skeleton estimation methods, OpenPose achieves the best lexeme assessment performance and performs comparably to some 3D methods. Importantly, the overall drop in F1 scores when moving from Kinect to RGB-based methods is not drastic, indicating the feasibility of webcam-compatible assessment systems. With respect to KL-HMM system configurations, the M configuration which models bilateral hand movement as a single unified channel consistently achieves the highest lexeme-level F1 scores. This suggests that, for assessing the correctness of a produced sign at the lexeme level, combining left and right hand movement into a single representation is more effective than modeling them separately.

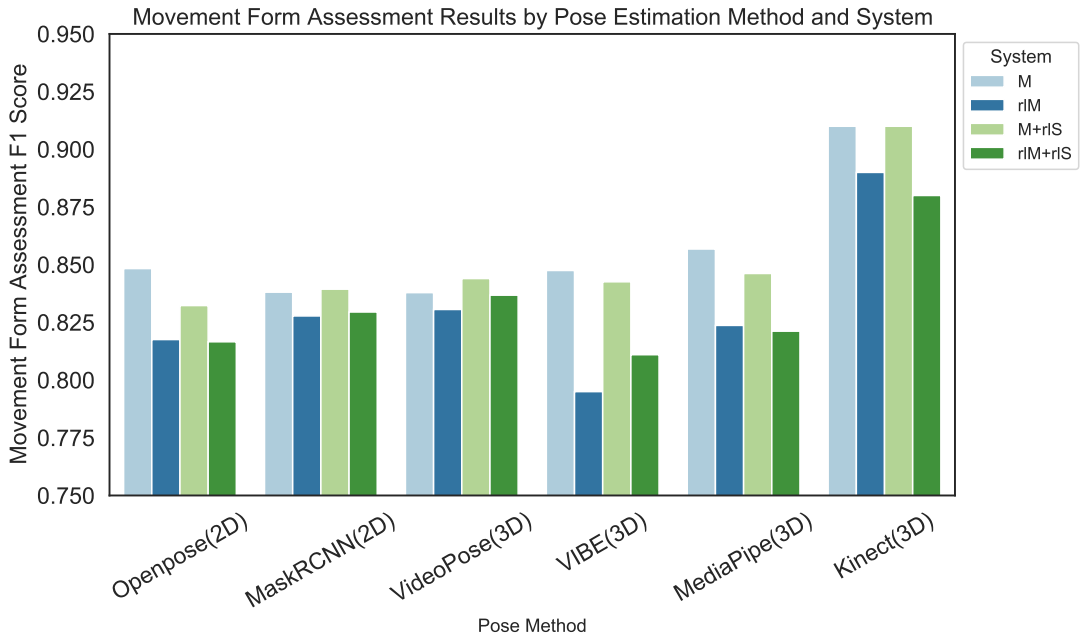


**Figure 3.5** – Movement form assessment F1 scores.

Figure 3.5 summarizes the hand movement form assessment F1 scores across skeleton estimation methods. Here, VideoPose3D achieves the highest scores in three out of the four KL-HMM configurations, confirming its effectiveness in capturing movement features relevant to form-level assessment. VIBE, once again, performs poorly in this context. Among the 2D methods, the Mask R-CNN achieves the best performance, outperforming OpenPose in 3 out of 4 configurations. These findings collectively indicate that accurate and accessible SLA both at the lexeme and form-level is achievable using RGB-based 3D skeleton estimation methods, particularly VideoPose3D and MediaPipe, which demonstrate strong capability in compensating for the lack of explicit depth information.

To analyse the statistical significance of the two best performing methods (VideoPose and Mask R-CNN), we conducted a McNemar's test on their lexeme and movement form assessment

outputs. The results indicate that the difference in lexeme assessment is not significantly different, whereas hand movement assessment is significantly different at 95%, owing to the differences in 2D and 3D movement features.

**Table 3.1** – Percentage of correctly identified movement assessment in signs across different directions of movement.

| **Pose Estimation** | **Movement Direction** | | | | |
|---|---|---|---|---|---|
| | $xy$ | $xyz$ | $z$ | **rotation** | **static** |
| Mask R-CNN | 66.99 | 51.94 | 61.32 | 64.13 | 50.00 |
| VideoPose3D | **78.47** | **74.42** | **73.58** | **81.52** | **59.09** |

**Figure 3.6** – Video snippets from the SMILE DSGS dataset showing sign productions illustrating different hand movement directions. The first row corresponds to the sign VON($z$), the second row corresponds to the sign BLAU($xyz$), and the third row corresponds to the sign EINVERSTANDEN($xy$).

To further examine the role of movement complexity, we categorized the test-set signs based on their dominant direction of hand movement. We identified five movement categories: (i) $x, y$ and $z$, (ii) $x, y$, (iii) $z$, (iv) in place wrist rotation, and (v) static sign. As illustrated in Figure 3.6, the first sign demonstrates movement along the $z$ axis, the second spans all three spatial dimensions, and the third is constrained to the $x$-$y$ plane. For each category, we compute the percentage of correctly assessed hand movements using both Mask R-CNN and VideoPose3D. Table 3.1 presents these results for the best-performing two-channel KL-HMM configuration (M+rlS). The data clearly show that VideoPose3D consistently outperforms Mask R-CNN across all movement types, particularly in cases involving depth-oriented motion. This supports the conclusion that 3D skeleton estimation offers tangible benefits for assessing hand movement accuracy in sign language.

To address RQ 3, we evaluated the same reference systems for isolated SLR, using them as classifiers rather than assessors. Figure 3.7 shows SLR accuracies for all skeleton estimation methods across different system configurations.

These results allow us to investigate the trade-off between SLR and SLA performance when transitioning from depth-based input to RGB-based skeleton estimation. As in assessment, the Kinect (3D) achieves the highest recognition accuracy across all configurations, confirming the benefit of true depth information. Among RGB-based methods, a notable observation is that the Mask R-CNN, despite being a 2D method, often outperforms VideoPose3D in terms

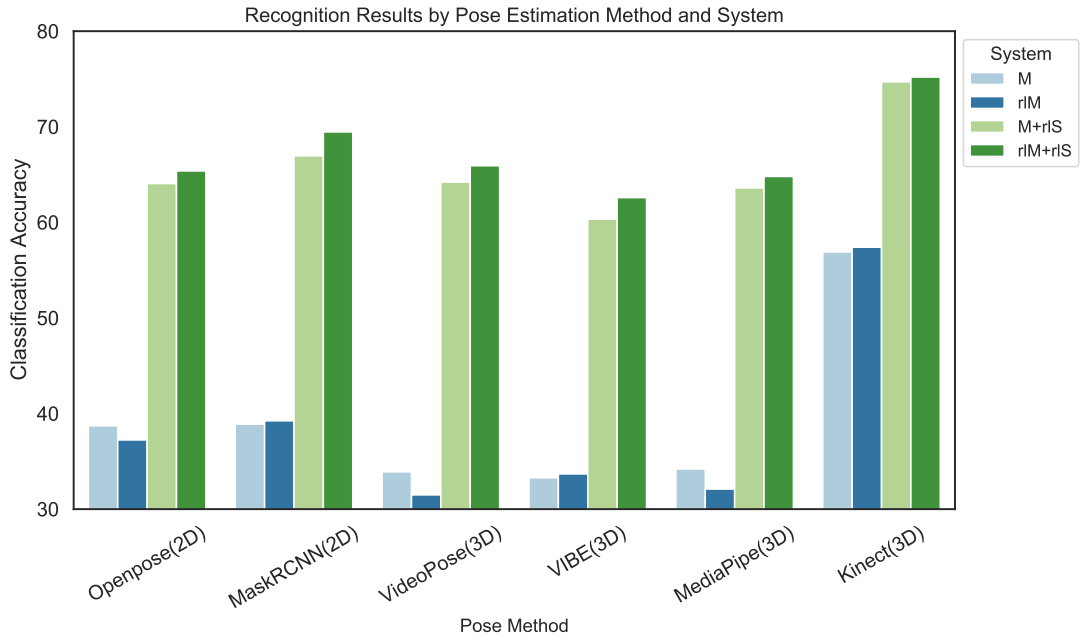


**Figure 3.7** – Sign language recognition accuracy of various skeleton estimation methods on the test dataset.

of recognition accuracy. This contrasts with the assessment result, where VideoPose3D consistently performed better. This discrepancy aligns with findings from (Arendsen et al., 2008), which reported that recognition performance does not necessarily correlate with acceptability or assessment ratings.

Overall, the degradation in recognition accuracy when moving from the Kinect to RGB-based methods is more pronounced than the drop observed in assessment performance. This suggests that while RGB-derived skeletons can retain enough structural information for evaluating sign correctness, they may be less effective for discriminating between different signs during classification, especially in the absence of reliable depth cues.

## 3.4 Summary

In this chapter, we investigated the feasibility of webcam-based accessible SLA using skeleton data derived from RGB videos, with the aim of replacing depth-sensor-based systems such as the Kinect. Through extensive experiments on the SMILE-DSGS dataset, we evaluated multiple skeleton estimation techniques: including 2D (OpenPose, Mask R-CNN) and RGB-based 3D methods (VideoPose3D, MediaPipe, VIBE) within a phonology-based assessment framework.

Our findings demonstrate that VideoPose3D and MediaPipe provide competitive performance relative to Kinect, particularly for lexeme-level and form-level assessment, making them strong candidates for webcam-compatible assessment systems. The Mask R-CNN, although

a 2D method, showed surprisingly strong performance in recognition tasks, whereas VIBE consistently underperformed across all configurations.

While skeleton-based approaches offer interpretability and modularity, they are limited by the accuracy and consistency of pose estimation models. In the following chapter, we shift focus to RGB-based holistic feature representations specifically, spatio-temporal embeddings extracted from raw video for SLA. This allows us to explore a complementary direction that bypasses pose estimation altogether and leverages deep visual features learned from large-scale pretrained models.

# 4 Deep learning-based representation for movement modeling

**Chapter Overview:** This chapter explores deep learning-based spatio-temporal representations for modeling hand movement in sign language assessment. We investigate and compare features extracted from I3D and Transformer-based architectures, analyzing their effectiveness within the assessment framework.

Until recently, most of the work in SLA has been in the framework of skeleton-based systems (Tornay, Camgoz, et al., 2020; Cory et al., 2024; Tarigopula et al., 2022; Arendsen et al., 2008), which rely on handcrafted features derived from estimated skeleton coordinates to model articulatory channels such as hand movement. While these approaches offer interpretability, particularly within the KL-HMM framework, which enables channel-wise modeling, they are highly sensitive to the quality of the skeleton extraction, like seen in Chapter 3. In addition to pose estimation errors, skeleton-based methods require explicit normalization steps to account for differences in body size, orientation, and camera perspective across videos. While such preprocessing is necessary to achieve consistent representations, it introduces further sensitivity to noise and variation. Moreover, handcrafted skeleton features fail to provide holistic representations of the signer's full visual and dynamic context, limiting their descriptive power for assessment tasks. In contrast, deep learning models are trained on large-scale datasets and possess significantly higher representational capacity, enabling them to learn complex spatio-temporal patterns directly from raw RGB input.

Despite their success in video classification tasks (Carreira et al., 2017; Arnab et al., 2021; Ryali et al., 2023) and their ability to encode powerful holistic representations, deep learning models face fundamental challenges when applied to SLA. Unlike recognition tasks, where the objective is to classify a sign or a sequence of signs, assessment requires a fine-grained analysis that identifies which components of a sign were produced incorrectly and explains why. A core limitation of many deep learning models is their lack of separability: although they effectively model spatio-temporal dynamics, they often entangle multiple articulatory features in their learned representations. This makes it difficult to attribute specific errors to particular sources such as handshape, movement, facial expression, etc. For sign language learning applications, where granular and channel-specific feedback is essential, this entanglement significantly

reduces interpretability and pedagogical value.

The goal of this chapter is to address these limitations by formulating approaches that combine the robust representation capabilities of deep learning models with statistical methods to enable fine-grained assessment of sign language videos. To address these limitations, we explore two complementary learning paradigms for extracting representations directly from RGB videos:

1. **Supervised task-specific learning**, where models are trained directly on annotated sign language data. This approach allows the learned representations to align closely with the visual and temporal patterns present in signing, enabling the system to capture domain-relevant features for assessment.

   This paradigm is instantiated using the I3D model for supervised learning, which is trained on MeineDGS[1] (Konrad et al., 2020), a large-scale dataset for German Sign Language (DGS), for the task of sign spotting. The results based on this work have been published previously in (Tarigopula et al., 2025).

2. **Use of self-supervised pre-trained representations**, where models trained on large-scale visual data using self-supervised learning objectives are employed as feature extractors. These models, although not fine-tuned on sign language data, provide rich and generalizable visual features that can be applied to assessment tasks in a zero-shot fashion.

   This paradigm is instantiated using transformer-based models such as DINOv2 (Oquab et al., 2023) and Hiera (Ryali et al., 2023) for self-supervised pre-trained representations and remain sign language agnostic.

The rest of the chapter is organized as follows. In Section 4.1, we describe the deep learning-based approaches explored in this work, detailing the architectures and feature representations used. Section 4.2 outlines the experimental setup, including the feature extraction, posterior feature conversion, and integration into the assessment framework. Section 4.3 presents an analysis of the I3D features, evaluating their separability and relevance for SLA. In Section 4.4, we report the quantitative results and compare the performance of the proposed deep learning-based methods to the skeleton-based baselines discussed in Chapter 3. Finally, Section 4.5 concludes the chapter with a summary of key findings.

## 4.1 Approaches

### 4.1.1 Inflated 3D ConvNets - I3D

The I3D (Two-Stream Inflated 3D ConvNet) model (Carreira et al., 2017) addresses the limitations of standard 3D convolutional networks, which are typically computationally expensive,

[1] DGS stands for Deutsch GebärdenSprache

relatively shallow, and unable to leverage powerful 2D pretraining (e.g., from ImageNet). I3D builds upon the Inception-v1 architecture, which exploits 1×1 convolutions for dimensionality reduction and feature fusion, by inflating its 2D convolutional and pooling kernels into 3D, effectively extending successful 2D image classification models into the spatio-temporal domain. This inflation process involves converting square $K \times K$ filters into cubic $K \times K \times K$ filters by adding a temporal dimension. To retain the benefits of ImageNet pretraining, the 2D weights are repeated along the temporal axis and normalized by the number of repetitions. This bootstrapping approach enables meaningful initialization of 3D filters and facilitates transfer learning from large-scale image datasets. As a result, I3D achieves strong performance on video-based tasks without requiring an entirely new 3D architecture. In this work, we use the single-stream I3D model based on RGB inputs to extract spatio-temporal features from sign language videos. These features are subsequently used for modeling articulatory dynamics within our assessment framework.

### 4.1.2 Transformers for vision

Vision Transformers (ViTs) (Dosovitskiy et al., 2021) represent a paradigm shift in visual representation learning by applying the self-attention mechanism, originally developed for natural language processing, to image data. Instead of processing the image as a whole, ViTs divide it into fixed-size patches (e.g., $16 \times 16$) which are then linearly embedded and treated as a sequence of tokens. These tokens are passed through a standard Transformer (Vaswani et al., 2017) encoder architecture, enabling the model to capture global context and long-range dependencies without relying on convolutional layers. Unlike CNNs, which have strong inductive biases for locality and translation invariance, ViTs learn spatial relationships more flexibly from data. When pretrained on large-scale datasets, ViTs achieve competitive or superior performance on many image recognition tasks. In our work, we explore ViT-based models for extracting high-level spatial representations relevant to hand movement and shape in sign language context.

**DINOv2** (Oquab et al., 2023) is one such model designed to learn robust and general-purpose visual representations in a self-supervised manner. It employs a student-teacher framework where a student network learns to replicate the representations of a momentum-updated teacher network. This approach enables the model to capture meaningful visual patterns without the need for labeled data.

**Hierarchical Vision Transformer** - Hiera (Ryali et al., 2023) is a simplified hierarchical ViT designed to achieve high accuracy and efficiency without relying on complex, vision-specific architectural components. Traditional hierarchical transformers (Li et al., 2022; Liu et al., 2021), often incorporate specialized modules like convolutions or shifted windows to introduce spatial biases, which can increase model complexity and reduce computational efficiency. In contrast, Hiera eliminates these components by leveraging a strong visual pretext task of Masked Auto Encoding during pretraining. The task involves masking random patches of the

input image and reconstructing the missing pixels. This approach enables the model to learn spatial biases directly from data, eliminating the need for added architectural complexity. It was also extended to video understanding by processing spatio-temporal patches and applying hierarchical attention across both space and time.

In summary, the I3D model is used within the supervised learning paradigm, where it is trained on annotated sign language data. In contrast, DINOv2 and Hiera are used within the self-supervised representation paradigm as frozen feature extractors, without any fine-tuning on sign language data.

## 4.2 Experimental setup

### 4.2.1 Feature extraction

#### I3D feature extraction

For modeling hand movement, we use the I3D model for action recognition, trained on the MeineDGS dataset (Konrad et al., 2020) for the task of sign spotting, i.e., recognizing isolated signs within specific frame windows (Tarigopula et al., 2025). While originally developed for action recognition, this model is expected to capture sign language-specific motion patterns due to its training on signing data.

Motivated by the structural similarities between sign languages, we apply a model trained on MeineDGS data to extract features for DSGS, aiming to assess its cross-lingual generalizability. The model takes 16 frames of size $224 \times 224$ as input, with necessary padding if signs last shorter than 16 frames. The model was trained to optimize cross-entropy loss using the SGD optimizer (Sutskever et al., 2013) with a momentum of 0.9, batch size of 4, and an initial learning rate of 0.01 with decay.

Feature extraction is performed using a sliding window approach, where each 16-frame segment is mapped to a 1024-dimensional feature vector from the penultimate layer of the I3D network. This feature is assigned to the central frame of the window, resulting in one representation per frame. Since the entire frame is processed, these features inherently encode both hand movement and handshape information. To better isolate movement-specific cues and reduce the influence of handshape in hand movement assessment, we additionally experiment with masking the hand region in the input frames prior to feature extraction. The framework for assessment using I3D features is illustrated in Figure 4.1.

#### ViT-based feature extraction

We use Vision Transformer (ViT) models—specifically DINOv2 and Hiera—to extract high-level frame-wise feature representations. For each video frame, we extract the embedding corresponding to the [CLS] token from the final transformer layer, which serves as a compact

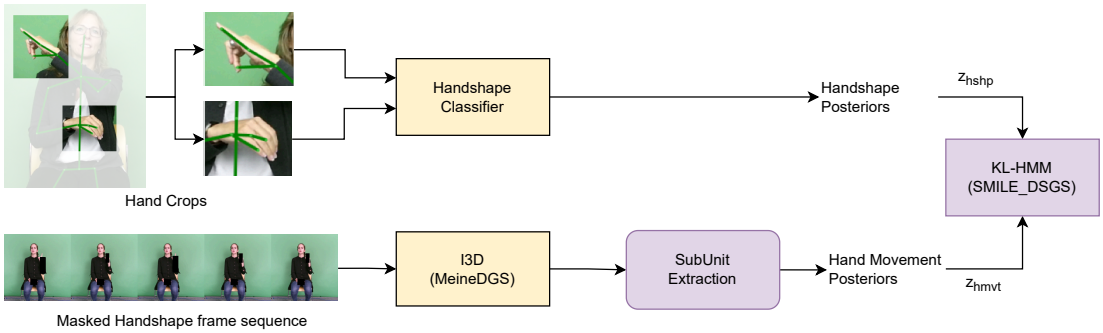


**Figure 4.1** – Framework for the development of deep-learning based systems for sign language assessment. Frame-wise hand crops are used to extract handshape posteriors, while hand masked sequences of 16 frames are used to extract hand movement posteriors. A stack of these posteriors is used to train the reference model.

global descriptor of the visual content.

DINOv2 is a self-supervised ViT model trained to produce semantically rich and context-aware representations that was trained on a curated dataset of 142 million images LVD-142M (Oquab et al., 2023). We experiment with two variants: ViT-Small, which produces 384-dimensional embeddings, and ViT-Base, which produces 768-dimensional embeddings.

In addition, we use the video variant of Hiera-B, a hierarchical Vision Transformer that processes spatio-temporal inputs and outputs a 768-dimensional [CLS] embedding. Thanks to its hierarchical architecture, Hiera is able to model both local motion patterns and long-range spatial-temporal dependencies. We use the version of the model that was pretrained on the Kinetics-400 (Kay et al., 2017) action recognition dataset.

### 4.2.2 Posterior feature conversion

To integrate the extracted deep features into the assessment framework, we convert them into frame-level posterior probabilities over a set of latent units. We explore two approaches for posteriorisation: direct probabilistic interpretation of feature activations using softmax normalization, and a subunit-based strategy that decomposes the representation space into multiple interpretable components.

#### Probabilistic interpretation of feature activations

Given a frame-level visual representation $\boldsymbol{f}_t \in \mathbb{R}^d$ extracted from a model, we first apply standard normalization:

$$\hat{\boldsymbol{f}}_t = \frac{\boldsymbol{f}_t - \mu}{\sigma}$$

where $\mu$ and $\sigma$ denote the mean and standard deviation, computed over the feature dimensions (globally). We then apply a softmax function to obtain a normalized posterior-like

representation:

$$\boldsymbol{z}_t = \text{softmax}(\hat{\boldsymbol{f}}_t), \quad \text{where} \quad z_{t,i} = \frac{e^{\hat{f}_{t,i}}}{\sum_{j=1}^{d} e^{f_{t,j}}}$$

The resulting vector $\boldsymbol{z}_t$ can be interpreted as a posterior distribution over the feature dimensions.

### Subunit-based posterior extraction

Subunit-based posterior extraction, introduced in Section 3.2.1, is also applied to the deep feature representations described in this chapter. As discussed previously, this approach involves modeling an articulatory channel using a set of latent subunits. For deep learning-based features, we follow the same procedure by first training an HMM/GMM system and then using the resulting state alignments to train an MLP classifier for posterior estimation. Importantly, this subunit modeling is trained on the SMILE-DSGS dataset, which may introduce language-specific characteristics into the resulting posterior distributions. While the pretrained feature extractor of I3D is trained on a different sign language data, the posterior conversion step is grounded on the target language data, i.e., DSGS.

We apply both posterior conversion methods to the I3D-based features. The subunit-based approach benefits from language-specific modeling, while the softmax-based approach offers a lightweight alternative without any language-specific modelling. For ViT-based features (DINOv2 and Hiera), we apply only the softmax-based method, as our goal is to examine the assessment pipeline using sign language agnostic features. This setup ensures that no sign language-specific supervision is introduced at any stage of feature extraction or posterior conversion, allowing us to evaluate the effectiveness of general-purpose visual representations for SLA.

### Handshape posterior extraction

As described earlier, we mask out the hand regions in the input frames used by the I3D model in order to suppress handshape-related visual cues and obtain representations that focus more selectively on hand movement. By comparing masked and unmasked I3D representations, we aim to assess the contribution of handshape information that might otherwise be entangled within the full-frame spatio-temporal features. To explicitly model handshape, we extract handshape posteriors independently using a CNN-based model trained on cropped hand regions. This model, introduced in Section 3.2.1, outputs frame-level posterior probabilities over a set of learned handshape units. These handshape posteriors are then integrated with either the masked or unmasked I3D movement features, enabling a controlled investigation of multimodal fusion strategies. This setup allows us to evaluate the additive effect of explicitly modeled handshape information when combined with spatio-temporal movement representations.

### 4.2.3 KL-HMM reference systems

As proposed in (Tornay, Camgoz, et al., 2020) for skeleton-based assessment, we adopt the phonology-based framework described in Section 2.2.1 for modeling the temporal structure of sign productions. Within this framework, each sign is modeled by a separate HMM trained on sequences of posterior features, with state emission probabilities defined over frame-level posterior distributions rather than raw features. This formulation allows us to directly compare a learner's posterior sequences against those of native signers using probabilistic divergence.

We train and evaluate different KL-HMM configurations based on the type of articulatory information included:

1. **M**: Models only the hand movement subunits, obtained from both the left and right hands (combined).
2. **M+S**: Models the concatenated features of hand movement subunits and handshape subunits.

For each sign class, KL-HMMs with varying numbers of states (ranging from 3 to 30) are trained, and the model yielding the best isolated sign recognition accuracy on the validation set is selected as the best model. In all cases, the reference models are trained using only **acceptable** sign productions — i.e., those labeled as Category 1 or 2 by expert annotators to ensure that the models reflect high-quality sign language productions.

For SLA, we follow the procedure outlined in Section 2.2.2 to determine decision thresholds for both lexeme-level and form-level evaluations (i.e., hand movement and handshape). These thresholds are derived using the validation set by computing KL divergence-based similarity scores: between pairs of acceptable sign instances for correct matches, and between instances of different signs for incorrect matches. For each assessment type, the threshold is selected to maximize the $F1$ score on the validation data, ensuring a balanced trade-off between precision and recall.

## 4.3 Discriminability analysis

A robust model for feature representations should be able to discriminate effectively between different sign classes while maintaining consistency within the same class. To better understand the quality and separability of the posterior representations derived from I3D features, we conduct a discriminability analysis comparing the two posterior extraction methods: the softmax-based approach and the subunit-based approach. We also include handcrafted skeleton-based subunit posteriors from (Tornay, Aran, et al., 2020) as a baseline.

For each method, we compute pairwise DTW distances between temporal sequences of posterior vectors. We distinguish between positive pairs (instances of the same sign class) and

negative pairs (instances of different sign classes). To account for different notions of similarity, we experiment with three distance measures as the DTW cost function: SKL divergence, KL divergence, and Bhattacharyya distance. This analysis is carried out only on training data of the SMILE DSGS dataset.

We analyze the resulting DTW distances across:

- **Positive pairs**: Sign instances belonging to the same sign class
- **Negative pairs**: Sign instances belonging to different sign classes

An ideal posterior representation should yield lower DTW distances for positive pairs and higher distances for negative pairs, thereby preserving lexeme-level structure in the feature space.

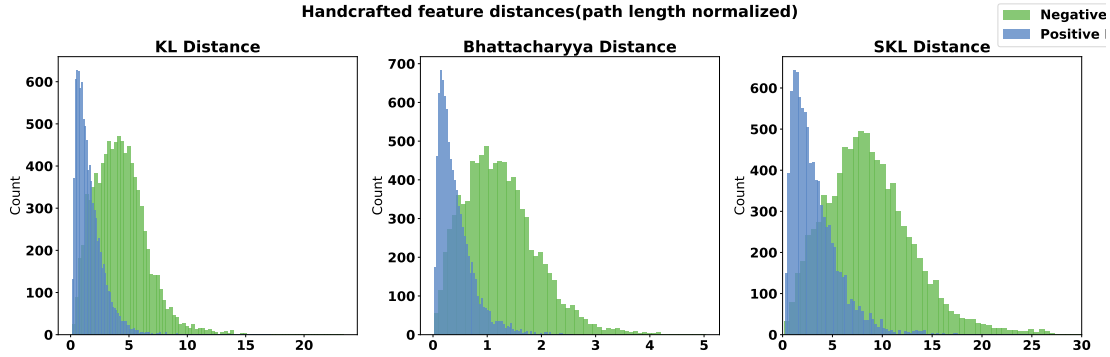


**Figure 4.2** – Histograms of DTW distances between sign pairs, comparing positive (same class) and negative (different class) pairs using handcrafted features.

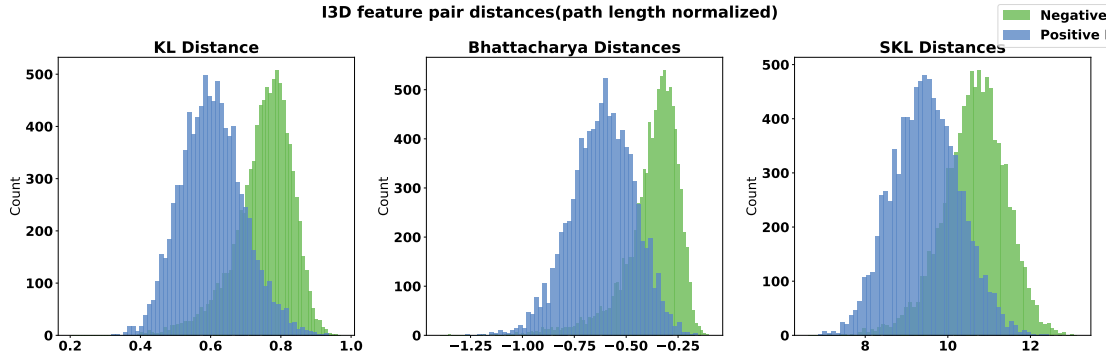


**Figure 4.3** – Histograms of DTW distances between sign pairs, comparing positive (same class) and negative (different class) pairs using subunit-based I3D posteriors.

We visualize the resulting distributions of path length normalized DTW distances using histograms in Figures 4.2, 4.3, and 4.4. The degree of overlap between the positive and negative distributions provides insight into the discriminative capacity of the features. Lower overlap

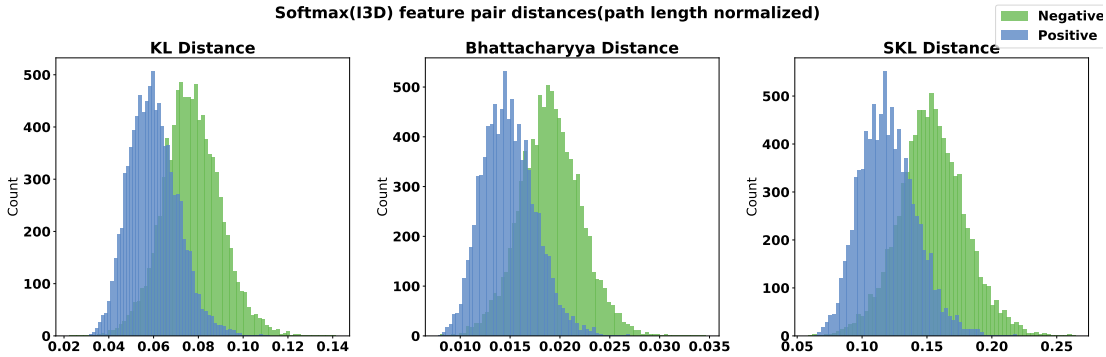


**Figure 4.4** – Histograms of DTW distances between sign pairs, comparing positive (same class) and negative (different class) pairs using softmax-based I3D posteriors.

indicates better class separability, suggesting that the features preserve sign-specific structure. In contrast, greater overlap implies limited ability to distinguish between different signs.

From the distance distribution plots, we observe that the I3D-based posteriors exhibit lower overlap between positive and negative pair distances. This suggests that they more effectively preserve sign-specific distinctions compared to the handcrafted feature based posterior used in earlier skeleton-based approaches. The reduced overlap suggests improved discriminative capacity, particularly at the lexeme-level.

These results highlight the suitability of subunit-based I3D posteriors for building robust reference systems in SLA. By leveraging language-specific alignment and classification, the subunit approach enables the extraction of temporally structured and class-separable I3D representations - both of which are critical for accurate and explainable assessment.

## 4.4 Results

In this section, we report recognition and assessment results obtained using different feature extraction and posterior conversion methods. For I3D-based features, we compare (i) softmax-based vs. subunit-based posterior representations and (ii) masked vs. unmasked inputs, where masking is used to suppress handshape cues and allow for explicit handshape modeling via fusion, and (iii) Comparisons with skeletal baselines. Results are reported for two modeling variants: **M** (hand movement only) and **M+S** (hand movement combined with handshape). This setup enables us to investigate the role of explicit handshape integration and the benefits of subunit modeling for assessment.

For ViT-based features, we analyze SLR accuracy across transformer layers and report assessment scores using only the final layer representations for the **M** variant. As these models are used without any sign language-specific training or fine-tuning, thus offering a view into the potential of sign language agnostic representations for SLA.

### 4.4.1 I3D-based

#### Recognition Results

To begin, we analyze the impact of the number of hidden states on SLR performance in the reference models. Figure 4.5 illustrates classification accuracies on the validation set for varying numbers of HMM states, ranging from 5 to 30. This analysis serves as a model selection step: the number of states yielding the highest SLR accuracy is selected for each configuration, and the corresponding model is then used for reporting test SLR and SLA results.

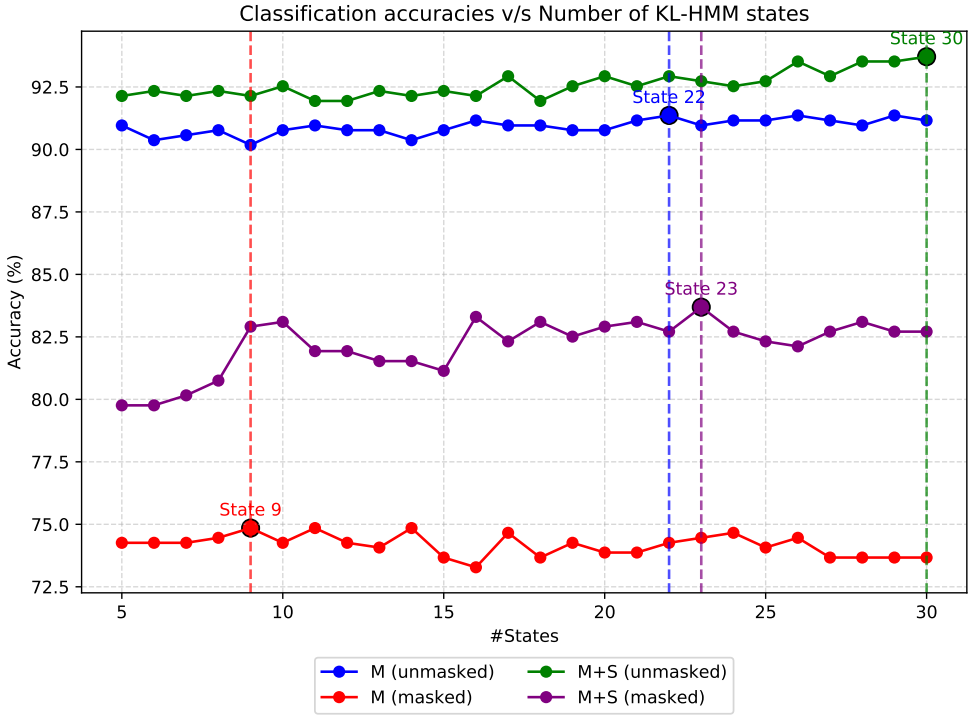


**Figure 4.5** – SLR accuracies for KL-HMM models with varying numbers of hidden states (5 to 30) on the validation set. The number of states yielding the highest SLR accuracy is selected for each configuration and used for final evaluation on the test set.

We evaluate the effect of posterior modeling strategy by comparing the softmax-based and subunit-based methods applied to unmasked I3D features. Both approaches are applied to unmasked I3D inputs, and results on the test data for SLR using the reference models are reported for the **M** (movement only) variant. As shown in Table 4.1, the subunit-based method outperforms the softmax-based approach in terms of recognition accuracy. This validates the advantage of structured, language informed modeling in producing more discriminative reference systems.

To benchmark the effectiveness of I3D-based spatio-temporal representations, we compare their SL recognition accuracy against the skeleton-based baselines introduced in Chapter 3. This comparison provides a reference point for evaluating whether deep RGB-based features

**Table 4.1** – SLR accuracy (%) for I3D-based features using subunit-based and softmax-based posterior extraction methods (**M** variant). The best performance is highlighted in bold.

| Posterior Method | Recognition Accuracy (%) |
|---|---|
| Subunit-based | **88.77** |
| Softmax-based | 83.74 |

when paired with KL-HMM modeling can outperform or complement traditional handcrafted skeleton features. For this comparison, we use the masked I3D configuration with subunit-based posterior extraction, which aligns most closely with the abstraction level of skeleton-based features. Masking the hand region in the input suppresses handshape and appearance-related information from movement, encouraging the model to focus on motion patterns analogous to what is captured by pose-based representations.

**Table 4.2** – SLR accuracy on the test set (%) for different model configurations. The best performance for each configuration is highlighted in bold.

| Feature Type | M | M+S |
|---|---|---|
| Skeleton (Kinect) (Tornay, Camgoz, et al., 2020) | 55.77 | 74.18 |
| Skeleton (VideoPose3D) (Tarigopula et al., 2022) | 33.90 | 65.92 |
| I3D-masked-subunit | **66.09** | **75.81** |

Table 4.2 summarizes the SLR accuracy on test set using skeleton-based and I3D-based hand movement features. The I3D-based models, using subunit posteriors from masked input, consistently outperform skeleton-based methods on the DSGS sign recognition task.

**Table 4.3** – SLR accuracy (%) using masked and unmasked I3D features for hand movement posterior extraction. The best performance for each configuration is highlighted in bold.

| Feature Type | M | M+S |
|---|---|---|
| I3D-unmasked-subunit | **88.77** | **89.65** |
| I3D-masked-subunit | 66.09 | 75.81 |

To further examine the impact of hand masking on hand movement feature extraction, we conduct additional experiments using unmasked I3D inputs. The recognition results for this configuration are presented in Table 4.3. The posteriors were extracted using the subunit method. In the unmasked setting, handshape information remains embedded within the spatio-temporal features, which, when combined with explicit handshape posteriors in the **M+S** variant, leads to improved recognition performance compared to the masked case. This suggests that retaining handshape cues in the visual input can be beneficial for SLR. However, its impact on assessment performance particularly in terms of interpretability and channel-wise error attribution remains unclear and is examined separately in the next section.

### Assessment Results

Table 4.4 presents the $F1$ scores for lexeme and form assessment for different systems. Higher $F1$ scores indicate a greater number of samples being accurately classified as correct or incorrect, in terms of lexeme and form assessment.

**Table 4.4** – Assessment performance in terms of $F1$ scores for handshape (**hshp**), hand movement (**hmvt**), and lexeme-level evaluation (**lexeme**). I3D-based models use subunit posterior extraction; hand masking is applied only in the I3D-masked configuration. The best performance is highlighted in bold.

| Model | Config | hshp | hmvt | lexeme |
|---|---|---|---|---|
| Skeleton (Kinect3D) (Tornay, Camgoz, et al., 2020) | M | – | 0.9003 | 0.8771 |
| | M+S | 0.7960 | 0.9049 | 0.8993 |
| Skeleton (VideoPose3D) (Tarigopula et al., 2022) | M | – | 0.8379 | 0.8427 |
| | M+S | 0.7993 | 0.8440 | 0.8644 |
| I3D-masked-subunit | M | – | *0.9222* | *0.9123* |
| | M+S | *0.8053* | *0.9090* | *0.9234* |
| I3D-unmasked-subunit | M | – | **0.9567** | **0.9532** |
| | M+S | **0.8065** | **0.9509** | **0.9473** |

In multi-channel systems (**M+S**), the best path is determined by combined scores across all articulatory channels. By comparing the performance of the combined system (**M+S**) to that of individual channels (**M**), we can analyze the impact of adding handshape information on both lexeme-level and channel-specific assessment. This comparison helps reveal whether additional information leads to complementary improvements or introduces interference in the fusion process. The following observations can be made from the results:

- **I3D-based models consistently outperform skeleton-based models** across all configurations, demonstrating the strength of deep spatio-temporal RGB representations for SLA.
- **In skeleton-based systems**, the **M+S** variant outperforms the **M** variant, indicating that explicitly modeling handshape complements hand movement features and improves overall assessment performance.
- **For the I3D-masked configuration**, adding handshape information in the **M+S** variant improves lexeme-level assessment but slightly degrades hand movement assessment, suggesting interference in the fusion process despite improved lexical performance.
- **For the I3D-unmasked configuration**, the **M+S** variant performs worse than **M** across lexeme and form-level assessments.
- Notably, handshape assessment scores remain similar between the masked and unmasked settings, indicating that handshape cues, when present in both input and as an

explicit channel, can influence the best path selection and negatively affect movement assessment in the **M+S** configuration.

### 4.4.2 ViT-based

In this section, we report the results of the ViT-based features for sign language assessment in a sign language-agnostic setting. The goal is to assess whether pretrained visual representations extracted from models such as DINOv2 and Hiera, which have not been exposed to any sign language data can support meaningful hand movement assessment when paired with KL-HMM modeling. We first analyze the SLR performance of the reference models using features extracted from different transformer layers. We then evaluate the assessment performance using $F1$ scores derived from the final-layer representations of each model, focusing on the **M** variant to isolate hand movement without additional modality-specific supervision.

#### Recognition

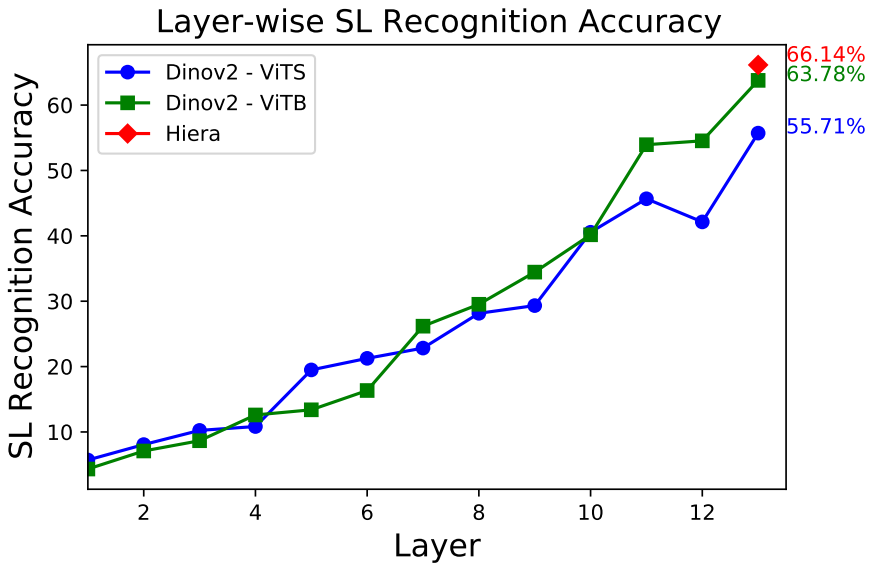


**Figure 4.6** – Layer-wise SL recognition accuracy using ViT-based features from DINOv2 (ViT-Small and ViT-Base) and Hiera. For each layer, frame-level features are extracted and used to train the reference models for sign classification.

Figure 4.6 presents the SLR accuracy using features extracted from different transformer layers of the DINOv2 ViT-Small and ViT-Base models. For each layer, frame-level features are converted into posterior distributions using the softmax-based method and used to train sign-specific reference models. For Hiera, recognition accuracy is reported only for the final layer due to its architectural design.

The results reveal a consistent trend for DINOv2: recognition accuracy improves in deeper

layers, indicating that higher layers encode more discriminative and task-relevant features. Among the two DINOv2 variants, ViT-Base outperforms ViT-Small, likely due to its larger representational capacity, which enables it to model more complex visual patterns. Notably, Hiera achieves the highest recognition accuracy overall, outperforming both ViT-Small and ViT-Base. This advantage can be attributed to Hiera's explicit spatio-temporal modeling, which allows it to capture motion dynamics across frames—an essential characteristic for sign language representation. Despite being trained without any sign language supervision, ViT-based models achieve recognition performance that is comparable to skeleton-based models in the **M+S** configuration, which explicitly incorporate both hand movement and handshape information and are trained on sign language data. This result highlights the strong representational capacity of pretrained ViTs and suggests their potential for SLR in low-resource settings, where sign language specific training data may be limited or unavailable. The corresponding assessment performance of these ViT-based features is studied in the next section.

#### Assessment

We now compare the form-level (hand movement) and lexeme-level assessment performance of ViT-based models with previously evaluated skeleton-based and I3D-based systems. All ViT-based features are extracted from the final transformer layer and converted to posterior representations using the softmax-based method, ensuring the entire pipeline remains sign language-agnostic. Only the **M** variant is considered in this evaluation to isolate the hand movement channel. The results are summarized in Table 4.5.

**Table 4.5** – $F1$ scores for hand movement (form-level) and lexeme-level assessment using different feature extraction methods. ViT-based models use softmax-based posterior extraction and are evaluated using the final layer. The overall best performance is highlighted in bold, while the best performance for ViT-based features is italicized.

| Model | Hand Movement | Lexeme |
|---|---|---|
| Skeleton (Kinect3D) (Tornay, Camgoz, et al., 2020) | 0.9003 | 0.8771 |
| Skeleton (VideoPose3D) (Tarigopula et al., 2022) | 0.8379 | 0.8427 |
| I3D-masked-subunit | 0.9222 | 0.9123 |
| I3D-unmasked-subunit | **0.9567** | **0.9532** |
| DINOv2 (ViT-Small) | 0.7484 | 0.7383 |
| DINOv2 (ViT-Base) | 0.7770 | *0.7414* |
| Hiera | *0.7818* | 0.7286 |

In terms of lexeme and hand movement assessment, the overall differences among ViT-based models are relatively modest. When compared to sign language aware models such as the skeleton-based systems or the I3D-based systems, all ViT-based models exhibit noticeably lower assessment performance, both in hand movement form and lexeme-level $F1$ scores. While ViT-based features achieve recognition accuracy comparable to skeleton-based **M+S**

models, their assessment performance is noticeably weaker. This discrepancy reinforces an important observation: high recognition accuracy does not necessarily imply strong assessment capability, particularly in contexts where detailed, component-wise evaluation of sign production is required.

## 4.5 Summary

In this chapter, we explored the use of deep learning-based representations for SLA, moving beyond the skeleton-based representations studied in the previous chapter. Our objective was to leverage the representational power of deep models while addressing the challenge of explainability and channel-wise assessment. We evaluated two learning paradigms: (i) a supervised model trained on sign language data (I3D), and (ii) pretrained self-supervised ViT-based models (DINOv2, Hiera) used in a sign language-agnostic manner. Features extracted from these models were converted into posterior representations using either subunit-based modeling or softmax normalization and integrated into the assessment framework.

Through a series of experiments, we showed that I3D-based features significantly outperform skeleton-based methods in both recognition and assessment tasks. We validate that language-specific information is effectively modeled through the integration of subunit extraction and reference model training on the target language. Interestingly, while masking the hand region allowed for more controlled separation of articulatory components, unmasked I3D features when paired with subunit modeling yielded the highest overall performance. However, adding explicit handshape features in the **M+S** variant sometimes led to degraded performance, highlighting the sensitivity of these features in the context of multi-channel fusion within KL-HMMs. ViT-based models, although effective in recognition, exhibited noticeably weaker assessment performance compared to sign language aware systems.

# 5 Towards sign language assessment in the loop for sign language generation

**Chapter Overview**: This chapter focuses on the use of phonology-based assessment methods to evaluate machine-generated sign language content. By comparing posterior feature sequences from reference and generated content, we aim to provide channel-wise objective scores that better reflect linguistic content than traditional quality metrics.

With the advancements in deep learning, SLG has emerged as a prominent area of research within sign language processing. The applications of SLG extend beyond communication and translation; they also include data augmentation for training recognition models, creation of educational content, and enhancing accessibility through synthesized sign language media. As previously noted in Chapter 2, SLG systems operate with a wide range of input modalities such as text, glosses, and phonetic representations, and they generate diverse outputs, including RGB video, 2D/3D skeletal poses, and avatar animations. However, as SLG systems become more widely adopted, it becomes crucial to evaluate not only the realism of the generated videos but also their usability, that is, whether they convey the intended linguistic message without introducing misleading information. Poorly generated signs risk propagating misinformation, reducing comprehensibility, or undermining learning, especially

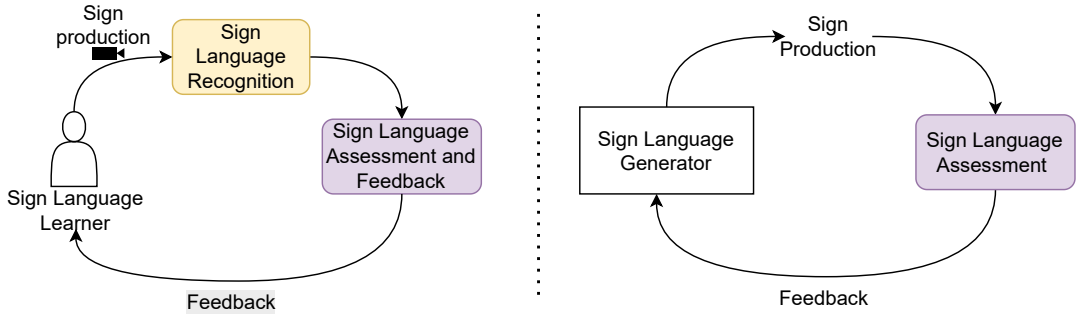


**Figure 5.1** – SLA for learners and SLG systems. The SLA framework can be extended to assess SLG outputs, providing feedback on articulatory quality and enabling a closed-loop system where assessment informs generation.

in educational settings.

SLA frameworks developed in previous chapters for learners can be naturally extended to SLG systems. While SLA has traditionally focused on providing interpretable feedback to human signers to help improve their performance, the same principles can be applied to assess SLG outputs at the frame level by evaluating articulatory quality. This concept is illustrated in Figure 5.1, where SLA serves not only as a diagnostic tool but also as a feedback mechanism for generation. In the long term, this enables the construction of a closed-loop system in which assessment and generation inform and enhance each other. In this chapter, we take a step toward that vision by developing assessment metrics that can serve as components of such a loop.

Reliable evaluation metrics are essential for measuring SLG quality. The use of automatic evaluation metrics is well established in machine translation (MT), where metrics like BLEU (Papineni et al., 2002) remain widely used despite newer metrics showing better alignment with human judgment (Freitag et al., 2022). Back-translation based evaluation uses such scores to evaluate generated content (Stoll et al., 2020; Vasani et al., 2020; Stoll et al., 2018). In the context where SLG outputs are visual and continuous in nature, conventional metrics like Peak-Signal-to-Noise-Ratio (PSNR) or Structural Similarity Index Measure (SSIM) are inadequate for capturing linguistic quality. These metrics are limited in their ability to capture the linguistic accuracy and articulatory clarity of generated signs, and may not fully align with human judgments.

To address this gap, there is a growing need for evaluation methods that are linguistically informed and better aligned with human judgment, ideally providing granular feedback on articulatory channels such as handshape, movement etc. Such methods would not only support more meaningful benchmarking of SLG systems but could also be integrated into generation pipelines to guide model improvement and ensure higher-quality, more interpretable outputs.

In this chapter, we propose the use of posterior-based metrics for evaluating reference-based SLG systems. We validate our approach on two SLG tasks:

- **video-to-video generation** in which sign language videos are synthesized from reference videos of another signer performing the same sign, in the context of video retargeting.

  For this generation task, we build upon the work of Krishna et al. (2021) to generate Swiss German Sign Language (DSGS) videos and evaluate them using both human judgments and automated metrics. The study, including a comparison with standard video metrics and our posterior-based approach, was published in (Tarigopula et al., 2024).

- **text-to-pose generation**, where gloss sequences are used to generate 2D pose representations of signing.

  For this generation task, we contribute to the work of Jiang, Leong, et al. (2024) by extending their pose-based evaluation study to include our proposed posterior-based

metrics. This collaboration, conducted as part of the Inclusive Information and Communication Technologies (IICT) project with the University of Zurich, aims to benchmark evaluation methods for sign language pose generation. By incorporating our posterior-based approach into this evaluation framework, we position it alongside a diverse set of existing metrics ranging from distance-based similarity to embedding-based distances and assess its alignment with human judgments. This comparative analysis not only validates the relevance of our proposed metric but also highlights its complementary role in the larger landscape of sign language evaluation tools.

The remainder of this chapter is organized as follows: Section 5.1 introduces the proposed posterior-based metrics for assessing the articulatory quality of generated sign language content. We then present two generation tasks: Section 5.2 details the video-to-video generation setup along with corresponding evaluation metrics and human evaluation protocol, followed by presentation of correlation studies. Section 5.3 describes the text-to-pose generation task, including the setup for evaluation metrics and human evaluation, along with correlation studies. Finally, we conclude with a summary of our findings in Section 5.4.

## 5.1 Posterior-based metrics

The framework for posterior-based matching is illustrated in Figure 5.2. The matching procedure follows the approach described in Section 2.2.2, with the key distinction that the KL-HMM state distributions are replaced by reference posteriors $[\boldsymbol{r}_{t'}^{hshp}, \ldots, \boldsymbol{r}_{t'}^{hmvt}]$ corresponding to the handshape and movement posteriors derived from the reference sign content. DTW is employed to align the generated sign posteriors to these reference posteriors, using a cost function based on the SKL divergence. The SKL scores calculated along the best path for each of the articulatory features $\text{SKL}_{hshp}$ and $\text{SKL}_{mvt}$ are used as metrics to assess the generated sign. A similar strategy was also explored in Chapter 7 of the thesis (Tornay, 2021) for single-view, reference-based SLA. Based on the output of the sign generator, movement posteriors can be extracted from the I3D model as described in Chapter 4 or using skeleton-based features as described in Chapter 3.

We choose the features depending on the content being generated. For the video-to-video generation task, we use both skeleton and I3D-based movement posteriors. For the text-to-pose generation task, we use only the skeleton-based features.

It is worth pointing out that the proposed approach is similar to the phone class conditional probability sequences based objective speech intelligibility assessment approach proposed in (Ullmann et al., 2015) for assessment of speech codecs and text-to-speech systems.

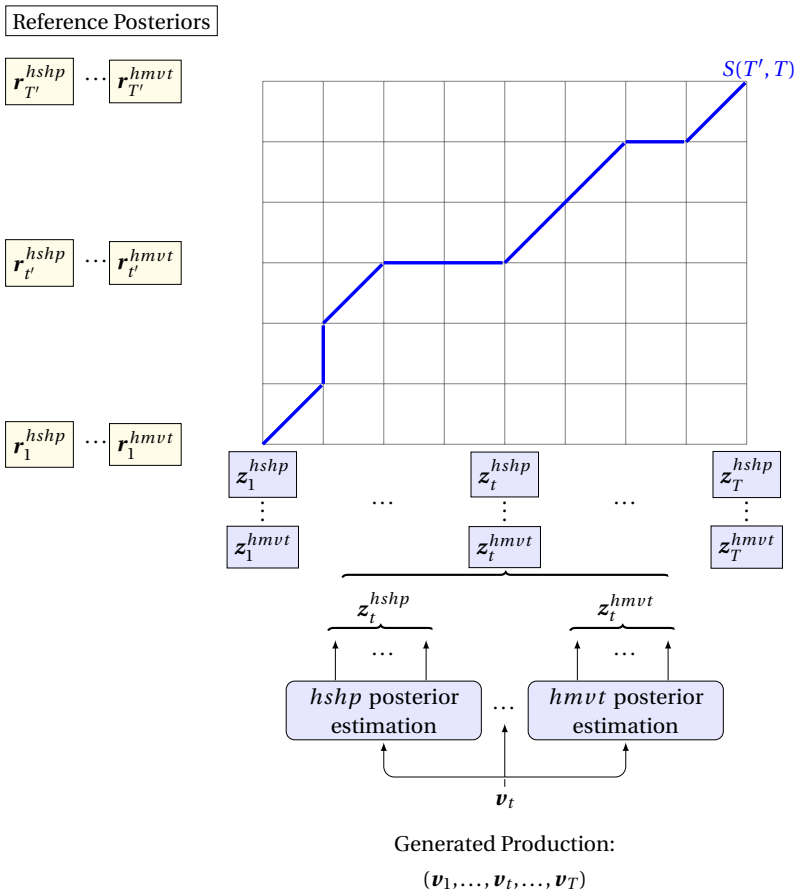


**Figure 5.2** – Illustration of posterior matching framework. $\left[z_t^{hshp}, \ldots, z_t^{hmvt}\right]$ is the stack of posterior estimates of the visual subunits obtained from the generated sign content $t$. $\left[r_{t'}^{hshp}, \ldots, r_{t'}^{hmvt}\right]$ is the stack of posterior estimates of the visual subunits obtained from the reference sign content. The DTW score is given by $S(T', T)$.

## 5.2 Video-to-video generation and assessment

We begin by presenting our study on the video-to-video generation task. In this section, we describe the generation method used to synthesize sign language videos from reference videos, outline the set of evaluation metrics employed including our proposed posterior-based metrics and present the results of their correlation with human evaluations.

### 5.2.1 Generation

Krishna et al. (2021) extend the Generative Adversarial Networks (GAN) based "do as I do" motion transfer model EBDN (Chan et al., 2019) for Indian SLG. Their approach employs two generators, one for the body and another specifically for the hands, to improve handshape generation. A smoothing network is used to seamlessly blend the outputs from the body and hand generators.

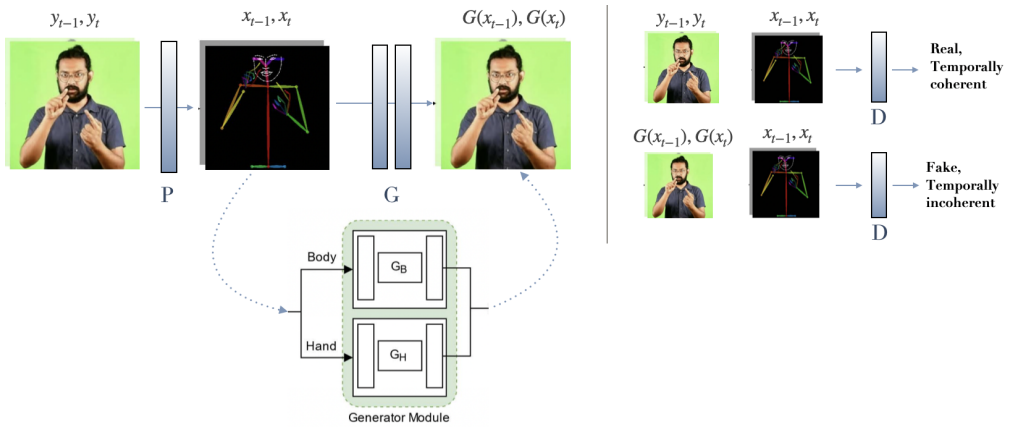


**Figure 5.3** – Training pipeline of Sign Language video generation from pose skeleton to images. Image adapted from (Krishna et al., 2021).

Given a video of a reference signer and another of a target signer, the model generates a video of the target signer performing the same action as the reference. An overview of the training pipeline is shown in Figure 5.3. The body generator $G_B$ generates upper-body images of the target signer conditioned on their pose skeleton, while the hand generator $G_H$ focuses on generating hand regions based on cropped hand skeletons. The pose estimator $P$ is used to extract these skeletal inputs. Both generators are based on the pix2pixHD framework (Isola et al., 2017), using a pair of networks operating at two spatial scales: global and local, denoted as $G = \{G_1, G_2\}$. Adversarial training involves three discriminators operating at different resolutions, $D = \{D_1, D_2, D_3\}$. To enforce temporal consistency, the model is trained to generate consecutive frame pairs, with a temporal discriminator distinguishing between real sequences $(x_{t-1}, x_t)$ and $(y_{t-1}, y_t)$ and generated sequences $(x_{t-1}, x_t)$ and $(G(x_{t-1}), G(x_t))$. Both $G_B$ and $G_H$ are trained using the same standard GAN loss function adapted for temporal consistency:

$$\begin{aligned}\mathcal{L}_{\text{temp}}(G, D) = \mathbb{E}_{(x,y)}\left[\log D(x_{t-1}, x_t, y_{t-1}, y_t)\right] \\ + \mathbb{E}_x\left[\log\left(1 - D(x_{t-1}, x_t, G(x_{t-1}), G(x_t))\right)\right],\end{aligned}$$

where $t$ denotes the time, $x$ refers to the input pose skeleton and $y$ to the real image.

The objective for GAN training is given by:

$$\begin{aligned}\min_G \Bigg( \max_{D_k} \Bigg( \sum_{k=1}^{3} \mathcal{L}_{\text{temp}}(G, D_k) \Bigg) + \lambda_{FM} \sum_{k=1}^{3} \mathcal{L}_{FM}(G, D_k) \\ + \lambda_p \mathcal{L}_p(G(x_t), y_t) + \lambda_p \mathcal{L}_p(G(x_{t-1}), y_{t-1}) \Bigg),\end{aligned}$$

where $\mathcal{L}_{FM}$ is the discriminator feature-matching loss (Isola et al., 2017) and $\mathcal{L}_p$ is the perceptual reconstruction loss that compares pre-trained VGG (Liu et al., 2015) features at different layers of the model.

We extend this work to generate signs for the DSGS sign language. Pose skeletons obtained from OpenPose (Cao et al., 2019) are used as input poses to the GAN, to generate the target person images. Figure 5.4 shows an example of a video generated for the DSGS sign "ABER".

**Figure 5.4** – First row shows the sequence of frames in the reference video for the sign "ABER", the second row shows the corresponding frames generated by the GAN.

### 5.2.2 Evaluation metrics

#### 5.2.2.1 Video quality-based metrics

We use the following baseline metrics to gauge the similarity between the reference video and the generated video:

1. **PSNR**: The ratio of the maximum possible value of the image and the power of noise that affects its quality (Bradski, 2000).

2. **SSIM**: Measures the similarity between two images based on their structural information, taking into account the strong inter-dependency among neighboring pixels. It produces a similarity score that reflects how closely the images match in terms of structure (Bradski, 2000).

3. **MSE Skeleton**: Mean-squared error (MSE) between the sequence of 3D skeletons from reference and generated videos obtained from MediaPipe (Lugaresi et al., 2019) inspired by Arkushin et al. (2023)

#### 5.2.2.2 Posterior-based metrics

To assess the quality of generated sign language videos beyond conventional image-level metrics, we propose the use of posterior-based metrics derived from linguistic subunits such as handshape and movement. The framework for obtaining the posterior-based metrics is shown in Figure 5.5. Given that we have the reference and generated videos, we can extract the corresponding hand movement posteriors using two approaches:

1. **I3D-based posteriors**: Hand movement features are extracted using a pretrained I3D model, and corresponding posteriors are obtained through subunit classification trained on the SMILE-DSGS dataset, as described in Chapter 4.

2. **Skeleton-based posteriors**: MediaPipe is used to extract skeleton features from the videos, and hand movement posteriors are computed via subunit classification, as detailed in Chapter 3.

For handshape posteriors, we leverage dynamic features extracted from hand skeletons using the method proposed by Tornay et al. (2025). The features consist of temporal trajectories of hand landmarks, specifically, the positions of finger joints relative to the wrist, capturing the dynamics of hand configurations over time. These relative coordinates help normalize for hand position and emphasize the articulatory shape. An HMM-GMM system is first trained to model these sequences and perform handshape classification based on their temporal structure. This system captures typical temporal patterns associated with different handshapes. The resulting state-level alignments from the HMM-GMM system are then used to train an MLP. The MLP is trained to map each feature observation to the state subunits. The final output is a frame-wise posterior distribution over handshape subunits, which can be integrated into the phonology-based assessment framework as a dedicated manual channel.

We use the posteriors obtained to match the reference and generated posterior sequences using DTW as described in Section 5.1 to extract the metrics: $\text{SKL}_{hshp}$ and $\text{SKL}_{mvt}$.

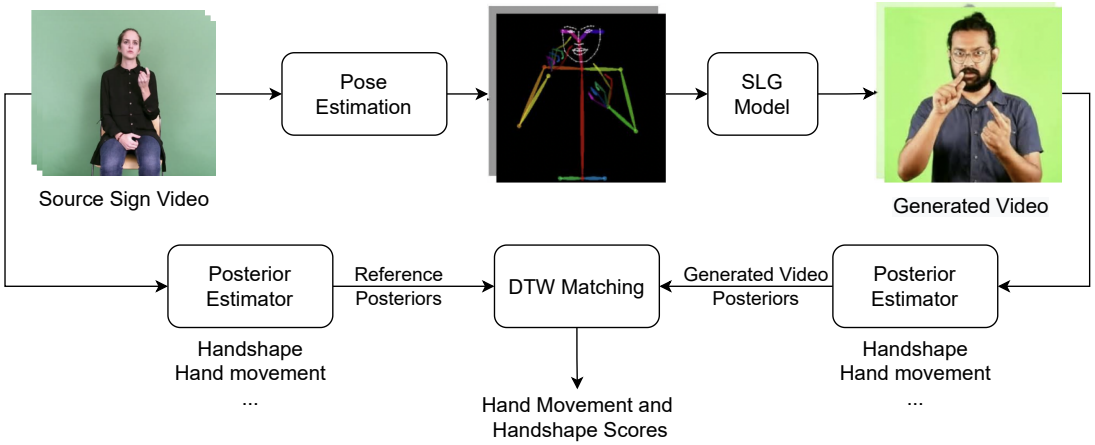


**Figure 5.5** – Framework for assessment of GAN-generated sign language videos

### 5.2.3 Human evaluation

For the subjective evaluation of the generated videos, 22 student raters were given 20 real/generated video pairs and asked to rate three questions for each video pair on a five-point Likert scale. The questions were designed to assess the quality of the generated video in terms of its similarity to the reference video, focusing on hand movement, handshape, and overall quality.

The questions were as follows:

1. How different is the hand movement in the generated video compared to the original video? (1:different - 5: similar)
2. How different is the handshape in the generated video compared to the original video? (1:different - 5: similar)
3. Rate the overall quality of the generated video (the higher the better).

**Table 5.1** – Intraclass Correlation Coefficient and statistics for rater agreement for each of the questions

| Question | ICC(3,k) | mean | std | skew |
|---|---|---|---|---|
| Q1 | 0.94 | 4.36 | 0.89 | -1.47 |
| Q2 | 0.96 | 4.0 | 0.93 | -0.84 |
| Q3 | 0.95 | 3.42 | 1.06 | -0.32 |

The Intraclass Correlation Coefficient (ICC) (Shrout et al., 1979) was used to get the agreement between the raters. A high ICC value (close to 1) indicates a high similarity between values. In our data, we observed high ICC values(ICC(3,k) > 0.90) indicating that all the raters had a high agreement and the excellent reliability of the raters. Table 5.1 summarizes the annotated questions, the ICC(3,k), and their respective descriptive statistics.

### 5.2.4 Correlation analysis

Table 5.2 presents Spearman's correlation coefficients ($\rho$) and p-values between automated evaluation metrics and human ratings across three aspects: movement ($HR_{mvt}$), handshape ($HR_{hshp}$), and overall sign quality ($HR_{all}$). Among conventional video quality metrics, SSIM shows moderate correlations with movement and high correlation with handshape. PSNR and MSE Skeleton exhibit weak or inconsistent correlations across aspects when compared to SSIM. Among the proposed posterior-based metrics, $\text{SKL}_{hshp}$ demonstrates strong correlations with handshape and overall quality, while $\text{SKL}_{mvt}$ (I3D) shows the best correlation with movement. The $\text{SKL}_{mvt}$ (MediaPipe) metric also exhibits moderate correlation with movement and overall quality. **The SKL scores are negatively correlated as low SKL corresponds to higher similarity, i.e., higher mean opinion scores**.

## 5.3 Text-to-pose generation and assessment

In this section, we shift our focus to the assessment of sign language pose sequences generated from text. The goal is to evaluate multiple evaluation metrics against human judgments, providing insights into the effectiveness of different metrics in capturing the quality of generated sign language content.

**Table 5.2** – Spearman's correlation coefficient and p-values between evaluation metrics and human ratings. $HR_{mvt}$, $HR_{mvt}$ and $HR_{mvt}$ denote human ratings for movement, handshape and overall quality, respectively. SKL$_{mvt}$ (MediaPipe) and SKL$_{mvt}$ (I3D) denote the posterior-based metrics for movement using MediaPipe and I3D, respectively. SKL$_{hshp}$ denotes the posterior-based metric for handshape. Statistically significant p-values are italicized.

| **Metric** | **Statistic** | **HR$_{mvt}$** | **HR$_{hshp}$** | **HR$_{all}$** |
|---|---|---|---|---|
| PSNR | corr | 0.18 | 0.067 | -0.085 |
| | p-value | 0.626 | 0.853 | 0.815 |
| SSIM | corr | **-0.51** | **-0.74** | -0.55 |
| | p-value | *0.041* | 0.084 | 0.190 |
| MSE Skeleton | corr | -0.54 | -0.42 | **-0.59** |
| | p-value | 0.106 | 0.228 | 0.069 |
| SKL$_{mvt}$ (Mediapipe) | corr | -0.58 | -0.58 | -0.45 |
| | p-value | *0.047* | *0.032* | *0.019* |
| SKL$_{hshp}$ | corr | NA | **-0.84** | **-0.79** |
| | p-value | NA | *0.002* | 0.006 |
| SKL$_{mvt}$ (I3D) | corr | **-0.61** | -0.49 | -0.52 |
| | p-value | *0.0244* | *0.0344* | 0.0849 |

### 5.3.1 Generation

We begin by describing the dataset and methods used to generate sign language pose sequences from text, forming the basis for our evaluation of different metrics. We consider three generation systems that reflect a range of approaches from rule-based retrieval to fully trainable neural models highlighting the diversity of strategies currently explored in text-to-sign language generation.

#### Dataset

For the text-to-pose task, the SignSuisse[1] dataset released as part of the WMT-SLT campaign was used. The dataset comprises 18,221 lexical items across three spoken–sign language pairs: German/Swiss German Sign Language (DSGS), French/French Sign Language (LSF), and Italian/Italian Sign Language (LIS). Each lexical item is represented by a signed example sentence presented in a video, accompanied by its corresponding spoken language translation. This forms a rich parallel corpus between sign and spoken languages, suitable for evaluating SLT and SLG models. For our experiments, we focus on the test set, which includes 500 DSGS segments, 250 LSF segments, and 250 LIS segments.

[1] https://www.sgb-fss.ch/signsuisse/

#### sign.mt

Based on the system proposed by Moryossef et al. (2023), this rule-based approach first converts input text into sign language glosses using hand-crafted transformations such as reordering and selective word deletion. Each gloss is then mapped to a corresponding sequence of skeletal poses retrieved from a predefined lexicon. The resulting poses are concatenated to form a complete sign sequence. When a gloss is not found in the lexicon, the system defaults to fingerspelling the word, generating pose sequences for each letter.

#### sign.mt v2

During initial evaluations, the frequent use of fingerspelling for unknown glosses was found to hinder the viewing and assessment experience for evaluators. To address this, an alternative version of the sign.mt system was introduced in which glosses with no lexical mappings are omitted entirely. While this may result in information loss, it significantly improves the fluency of the generated sequence and reduces cognitive load on evaluators.

#### Sockeye

The neural machine translation framework proposed by Hieber et al. (2022) was modified to generate continuous sign pose sequences. Both the encoder and decoder were modified to handle continuous output sequences instead of discrete tokens. In particular, rather than using a discrete vocabulary with an embedding matrix, the continuous pose vectors are linearly projected into the model's hidden dimension to serve as input to the decoder. This allows the model to directly learn mappings between text inputs and continuous articulatory pose outputs. The model was trained to predict pose sequences from text sequences from the SignSuisse training set.

The SignSuisse data is used to evaluate the performance of different text-to-pose translation systems. In the following section, we describe the set of metrics used in the evaluation

### 5.3.2 Pose-based evaluation metrics

Figure 5.6 shows the taxonomy of the evaluation metrics used in this study.

#### 5.3.2.1 Pose distance-based metrics

Prior work, such as Ham2Pose (Arkushin et al., 2023), evaluates sign language poses using distance-based metrics like Mean Joint Error (MJE) and Average Position Error (APE), with variants that apply normalization, alignment, or different aggregation strategies. One such approach uses DTW matching, comparing only shared keypoints between poses and penalizing keypoints present in the reference but missing in the generated pose.

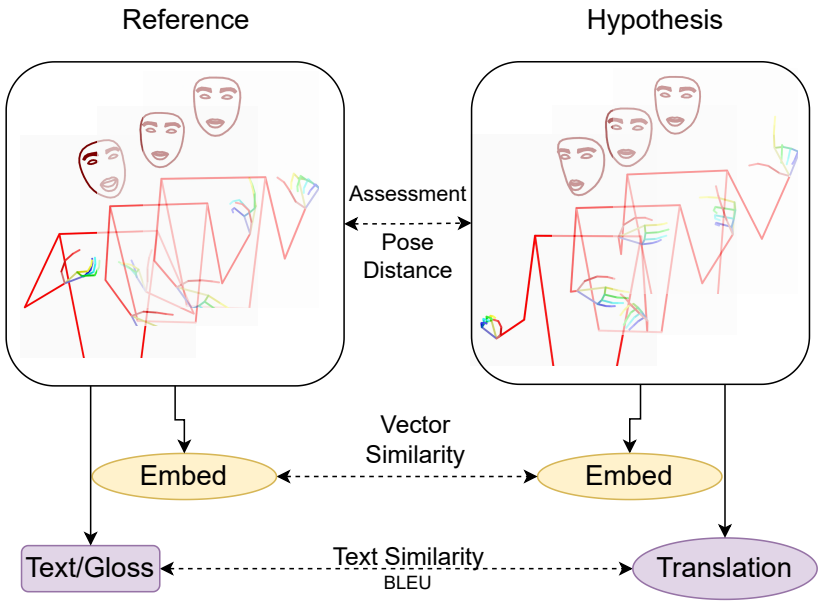


**Figure 5.6** – Taxonomy of pose-based sign language evaluation metrics.

These metrics originate from pose estimation and reconstruction literature and are commonly used in time-series clustering to compare trajectory shapes (Aghabozorgi et al., 2015). However, these methods are not inherently designed for sign language, and their naïve application can overlook critical issues: they assume complete keypoint detection, ignore performance speed variations, and conflate positional differences with semantic differences. Furthermore, they have not been validated on sign language data correlating with human judgments.

#### 5.3.2.2 SLA-based metrics

Two SLA-inspired metrics are considered for evaluating the quality of generated pose sequences: the SkeletonVAE Score, proposed in (Cory et al., 2024), and our posterior-based SKL Score, which uses linguistically informed features.

**SkeletonVAE score**

The SkeletonVAE model is trained to produce a per-frame latent embedding from 3D skeletal poses. To obtain these, 2D MediaPipe poses are first uplifted to constrained 3D skeletons using the method of Ivashechkin et al. (2023), and then projected into a 10-dimensional latent space using a $\beta$-VAE (Higgins et al., 2017). The SkeletonVAE Score is defined as the L2 distance between the DTW-aligned latent trajectories of the reference and hypothesis sequences, optionally normalized by the DTW path length. This metric captures pose similarity in a learned, compact representation space of sign language poses.

#### Skeleton posterior-based SKL score

As discussed in Section 5.1, two sets of linguistically informed posteriors—corresponding to handshape and hand movement—are extracted from the pose sequences, using the same missing keypoint preprocessing strategy as in Arkushin et al. (2023).

For hand movement, we use a subunit MLP trained on DSGS with MediaPipe skeleton data to compute movement posteriors for both the reference and hypothesis sequences. Similarly, for handshape, we use a separate subunit MLP, also trained on DSGS, to compute handshape posteriors from the corresponding pose sequences. It is important to note that these MLPs are trained on a different sign language (DSGS) than the target sign language being evaluated, which may introduce cross-lingual variation but still provide meaningful articulatory signals.

The resulting handshape and movement posterior sequences from the reference and hypothesis are then aligned using DTW, with a cost function based on the SKL divergence. The total alignment cost is aggregated over all DTW time steps to produce the final evaluation score, reported as $SKL_{mvt\_hshp}$ Score, and is used as the metric for the evaluation.

### 5.3.2.3 Embedding-based metrics

Extending the idea of *CLIPScore* (Hessel et al., 2021), SignCLIPScore (Jiang, Sant, et al., 2024) is an embedding-based metric designed specifically for sign language pose sequences. It uses SignCLIP, a model trained through multilingual contrastive learning to encode sign language poses into a shared semantic space. The metric defined as *SignCLIPScore P-P* (pose-to-pose) based on the dot product of the embeddings of the reference and hypothesis on the sentence level instead of frame-level latents plus DTW alignment. Another variant called *SignCLIPScore P-T* (pose-to-text) was also used. It computes the dot product between the text and pose embedding, eliminating reliance on scarce or even unreliable ground-truth signing references.

### 5.3.2.4 Back-translation-based metrics

Assuming the availability of corresponding spoken language text and a reliable pose-to-text SLT model, sign language pose sequences can be evaluated through back translation using two strategies:

(a) **Translation**: The pose sequence is translated into text using an SLT model, and the resulting output is compared to the reference text using standard machine translation metrics such as

1. *BLEU*: A precision-based metric that measures word-level n-gram overlap between the generated and reference.

2. *chrF*: A character n-gram F-score that captures fine-grained similarity, which is robust to minor lexical or morphological variation.

3. *BLEURT*: A learned evaluation metric based on pretrained language models, designed to align closely with human judgments of semantic quality.

(b) **Likelihood**: The log-likelihood of the reference text is computed given the pose sequence as input to the SLT model. This approach avoids decoding errors and supports more consistent system-level comparisons. Both methods depend heavily on the quality of the SLT model. For evaluation in this study, an SLT model from Zhang et al., 2024 is used.

All these automatic metrics are computed on the generated pose sequences across the three systems and compared against human evaluations. In the next section, we describe the human evaluation protocol used in this study.

### 5.3.3 Human evaluation

Evaluators rate the translations on a continuous scale from 0 to 100, with 0–6 markers and task-specific annotation guidelines adapted from WMT-SLT. Instructions are provided in DSGS, LSF, and LIS. A total of 13 native deaf signers participated (7 DSGS, 2 LSF, 4 LIS), with varying professional experience as translators, interpreters, or teachers. The evaluation covered 2,650 unique examples, yielding 11,471 ratings across four systems and three sign languages. The inter-annotator agreement, approximated using Fleiss' $\kappa$ (Fleiss, 1971) on discretized scores, was $\kappa = 0.36 \pm 0.05$, while intra-annotator agreement averaged $\kappa = 0.49 \pm 0.09$. The low inter- and intra-annotator agreement can be attributed to the absence of clear definitions and criteria for evaluating translation quality in signing poses.

### 5.3.4 Correlation analysis

Table 5.3 shows the correlation analysis between the metrics described earlier, divided into different families and the human scores averaged over evaluators on the segment level, as presented in Section 5.3.3. The correlation is computed using Spearman's rank correlation coefficient. **The signs of the metrics that quantify errors are flipped to keep a positive correlation for analytical convenience.** The following observations can be made from the correlation analysis:

- **Distance-based metrics**: These are effective when properly tuned, but highly sensitive to implementation details like pose format and keypoint selection.
- **SLA metrics**: SkeletonVAE scores show moderate correlation with human ratings at the language-level (where the scores across different systems for a language are evaluated together), but are less reliable for system-level evaluation in open-ended translation tasks. The SKL scores show stronger correlations at the system-level compared to other metrics. Interestingly, at the language level, SKL scores show strong negative correlations with human ratings, in contrast to the positive correlations observed with other metrics.

**Table 5.3** – Segment-level Spearman correlations with average human judgments calculated for several pose-based evaluation metrics for sign language. nAPE=`normalized APE`, nDTW=`normalized DTW-MJE` (two metrics taken from Arkushin et al. (2023) and re-implemented for MediaPipe, normalized by shoulder pose); SVAE=`SkeletonVAE Score`, $SVAE_n$=`SVAE normalized by DTW path`, SKL=`SKL_mvt Score`; P-P=`Pose-to-pose embedding distance`, P-T=`Pose-to-text embedding distance`; B4=`BLEU-4`, chrF=`chrF`, B-RT=`BLEURT`, Lik. =`Likelihood`. H* denotes mean inter-evaluator Spearman correlation. SD represents the standard deviation across each column and is expected to be small/consistent for an ideal metric.

| | **Reference-Based** | | | | | **Reference-Free** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Distance-Based** | | **SLA Metrics** | | | **SignCLIPScore** | | **Back Translation-Based** | | | | **H*** |
| | **nAPE** | **nDTW** | **SVAE** | **$SVAE_n$** | **SKL** | **P-P** | **P-T** | **B4** | **chrF** | **B-RT** | **Lik.** | **H*** |
| *By System* | | | | | | | | | | | | |
| sign.mt | 0.09 | 0.14 | 0.23 | -0.08 | 0.24 | 0.10 | 0.02 | 0.05 | 0.11 | 0.05 | 0.23 | 0.43 |
| sign.mt v2 | 0.28 | 0.33 | 0.46 | 0.14 | 0.22 | 0.00 | -0.19 | 0.20 | 0.22 | 0.44 | 0.49 | 0.52 |
| Sockeye | 0.10 | 0.15 | 0.13 | 0.01 | 0.24 | 0.42 | -0.27 | -0.07 | 0.04 | 0.46 | 0.58 | 0.22 |
| *By Language* | | | | | | | | | | | | |
| DE→DSGS | -0.36 | -0.09 | -0.02 | 0.27 | -0.57 | -0.31 | 0.39 | 0.18 | 0.26 | 0.09 | 0.36 | 0.70 |
| FR→LSF | -0.54 | -0.11 | -0.01 | 0.37 | -0.68 | -0.01 | 0.45 | 0.32 | 0.60 | 0.47 | 0.29 | 0.80 |
| IT→LIS | -0.57 | -0.39 | -0.02 | 0.53 | -0.75 | 0.13 | 0.29 | 0.31 | 0.63 | 0.41 | 0.38 | 0.88 |
| Overall (↑) | -0.41 | -0.10 | 0.07 | 0.38 | -0.56 | -0.10 | 0.27 | 0.21 | 0.42 | 0.36 | 0.42 | 0.77 |
| SD (↓) | (0.35) | (0.24) | (0.18) | (0.22) | (0.47) | (0.22) | (0.29) | (0.14) | (0.23) | (0.18) | (0.12) | (0.24) |

Upon further analysis, we observed that the SKL scores for the sign.mt v2 system were consistently higher on average than those of the other systems. The higher scores for sign.mt v2 are likely due to frequent missing keypoints in the system, which disrupt DTW alignment and lead to higher overall scores. Due to the lack of a common scale of scores across systems, in a setting where scores from multiple systems are compared jointly, the SKL score may not be the best metric to use.

- **Embedding-based metrics (SignCLIP)**: The pose-to-pose metric shows no meaningful correlation, whereas the pose-to-metric at language-level shows a moderate correlation.
- **Back translation metrics**: These align reasonably with human judgments and are valuable in general settings, though still fall short of human-level agreement.
- **Back translation likelihood**: These are more consistent and reliable than text-based metrics (BLEU, chrF, BLEURT) when using a pose-to-text model.

## 5.4 Summary

In this chapter, we explored the use of posterior-based metrics for evaluating reference-based SLG systems, with a focus on assessing the linguistic quality and intelligibility of generated content. Unlike conventional video metrics such as PSNR or SSIM, which primarily measure low-level visual similarity, our proposed metrics aim to capture articulatory correctness by analyzing posterior probabilities from SLR models.

We evaluated the effectiveness of these metrics in two SLG scenarios. In the video-to-video generation task, we built upon existing work in sign video retargeting for DSGS and conducted both human and automatic evaluations. Our posterior-based metrics demonstrated stronger alignment with human ratings than traditional video quality metrics. In the text-to-pose generation task, we extended an existing pose evaluation framework to include our metric. We observed a moderate correlation with human judgments at the system level, and a strong negative correlation at the language level.

# 6 Non-manual feature detection for continuous sign language

**Chapter Overview**: This chapter takes a step toward non-manual feature detection in the context of continuous sign language by modeling and evaluating facial action units (FAU) as linguistic cues. We explore FAU-based detection methods for non-manual feature detection and integrate them into a phonological framework to assess their impact on sign segmentation.

Continuous sign language (CSL) consists of a sequence of signs produced without clear pauses or boundaries, making it a sequence-to-sequence phenomenon. As mentioned earlier, unlike gesture recognition, sign language relies on multiple articulatory channels manual (e.g., handshape, movement, location) and non-manual (e.g., facial expressions, head movements, and body posture) to convey information. This multi-channel nature, coupled with phenomena like co-articulation (transitions between consecutive signs), makes segmentation in CSL particularly challenging (Cooper et al., 2012).

Non-manual markers play a crucial role in sign language grammar and semantics (Mukushev et al., 2020). Examples include:

- Raised eyebrows to mark Wh-questions (e.g., “what,” “where”)
- Forward or backward body movements to indicate tense
- Cheek puffing or retraction to express quantity
- Lip shape or motion to differentiate signs sharing the same manual form
- Facial expressions to convey emotional content
- Eye gaze plays a role along with a pointing index finger to refer to an object/person
- Head tilts for affirmation or negation

These features are especially critical in continuous signing, where sentence-level structures rely on non-manual cues. Effectively incorporating non-manual information is therefore essential for both understanding and assessing sign language. *We define non-manual assessment as the detection of the presence or absence of each non-manual feature in comparison to a reference, such as expert-annotated ground truth or target production.*

Prior work has explored various forms of non-manual feature modeling strategies, including mouthing cues (Albanie et al., 2020; Hu et al., 2021), 3D face scans (Kratimenos et al., 2021), and 2D facial landmarks (Asteriadis et al., 2012). While these methods capture certain aspects of facial expression or emotion, they often lack the granularity needed for fine-grained assessment. Additionally, they are typically embedded within end-to-end recognition models, making it difficult to extract explainable, frame-level probabilistic cues suitable for integration into assessment frameworks. Explicit detection and modeling of non-manual features remains under-researched, largely due to the limited availability of annotated datasets with detailed non-manual information.

In contrast, the Facial Action Coding System (FACS) (Rothkrantz et al., 2009; Ekman et al., 1978) provides a structured representation of facial expressions through FAUs, which correspond to localized muscle activations (e.g., cheek raise, puff, brow lift). FAU detection has been applied in sign language processing (Viegas et al., 2023; Silva et al., 2022) and aligns well with our objective of generating frame-wise posterior probabilities for interpretable non-manual behaviors. Therefore, we adopt FAU detection-based non-manual feature (NMF) modeling in this chapter. However, FAUs do not fully align with sign language-specific non-manual categories (see Section 2.3.3). To bridge this gap, we propose to fine-tune FAU-trained models on the SMILE-SRT dataset, which contains annotated non-manual features specific to sign language in continuous settings.

This chapter presents our initial investigation into non-manual feature detection and their integration into a phonology-based CSL framework. Specifically, we assess the contribution of non-manual features to temporal sign segmentation by comparing alignment performance with and without non-manual cues. Alignment refers to the process of mapping a sequence of observations (e.g., video frames or acoustic frames) to a sequence of model states or units (e.g., signs, phonemes, or subunits). Our approach leverages the KL-HMM framework, which supports multi-channel input and enables temporal alignment of signs in continuous data.

The remainder of this chapter is organized as follows: Section 6.1 discusses the different approaches we use for non-manual feature detection through Facial Action Unit detection. Section 6.2 details the integration of these features into the phonology-based framework and presents results from our alignment experiments. Section 6.3 concludes the chapter with a summary.

## 6.1 Non-manual feature detection

### 6.1.1 Detection methods

In this chapter we discuss the different approaches we use for NMF detection through FAU detection.

#### Face detection

Face detection is a crucial first step in NMF detection, as it enables the system to localize the face region and extract relevant features. In this chapter, we employ the Multi-Task Convolutional Neural Network (MTCNN) (Xiang et al., 2017) face detector to identify and crop the face region from individual video frames. For frames where a face is not detected, we apply interpolation using neighboring successful detections to maintain temporal consistency in video-based approaches. MTCNN uses a cascade of three CNNs to progressively refine face proposals and predict facial landmarks. This architecture provides a robust and efficient solution for face detection under varying lighting, pose, and occlusion conditions.

#### FLAME-based

Faces Learned with an Articulated Model and Expressions (FLAME) (Li et al., 2017) is a statistical 3D model of the face with parameters for identity shape, facial expression and pose parameters. In our work, we propose to use the facial expression-based parameters from this framework for non-manual feature detection. The FLAME coefficients are computed from the video frames using a pre-trained model, which allows us to capture the facial expressions in a continuous manner.

Emotion Driven Monocular Face Capture and Animation (EMOCA) (Danecek et al., 2022; Feng et al., 2021) introduces a method for reconstructing expressive 3D facial geometry from single, in-the-wild RGB images. It leverages the FLAME parametric face model to accurately capture detailed facial expressions that reflect the emotional state of the subject. The primary motivation behind EMOCA was to address the limitation of prior methods in accurately recovering 3D facial shapes from single images across the full emotional spectrum. In our approach, we utilize EMOCA's single-view fitter to extract the FLAME coefficients from video frames, ensuring robust and detailed facial expression representation.

#### I3D-based

As discussed in Chapter 4, Inflated 3D Convolutional Networks (I3D) effectively capture spatio-temporal representations from videos. Facial expressions typically exhibit dynamic patterns involving distinct onset, peak, and offset phases. Certain transitions between expressions can be subtle and thus require analysis across multiple frames rather than relying solely on

individual frames. Temporal modeling approaches using LSTM-based models have been successfully applied to emotion recognition tasks to address this challenge (Pyumina et al., 2022; Corneanu et al., 2018; Tellamekala et al., 2024). However, due to the high cost associated with labeling FAUs, large-scale video datasets with frame-level annotations remain relatively scarce. One notable exception is the Aff-Wild2 dataset (Kollias et al., 2023b), as described in Chapter 2. In this chapter, we leverage an I3D model to extract spatio-temporal features from facial video frames, which we subsequently utilize for non-manual feature detection.

#### Transformer-based

Vision Transformers (ViTs) (Oquab et al., 2023; Ryali et al., 2023; Liu et al., 2021) have shown promising results in a variety of computer vision tasks, and are typically pre-trained on large-scale datasets such as ImageNet. However, in most natural images, the face region occupies only a small portion of the image. As a result, standard ViTs often struggle to capture fine-grained facial representations effectively (Dan et al., 2023). To overcome this limitation, transformer-based models for facial analysis need to be trained on large-scale face-centric datasets or use specialized data augmentations to improve performance (Ma et al., 2023; Zhong et al., 2021; Ning et al., 2025). One such method proposes a Facial Masked AutoEncoder (FMAE) (Ning et al., 2025), an approach specifically designed for facial representation learning. By pretraining on large-scale facial datasets, FMAE learns robust and expressive facial features, which are beneficial for downstream tasks such as FAU detection and subsequent non-manual feature analysis in sign language understanding.

### 6.1.2 Experimental setup

#### FLAME-based FAU detection

In this approach, we utilize FLAME expression coefficients as a compact, interpretable representation of facial expressions. For each frame in the dataset, we first apply MTCNN (Xiang et al., 2017) to detect and crop the face region. The cropped face is then passed through EMOCA to extract the FLAME expression coefficients, resulting in a 30-dimensional vector per frame.

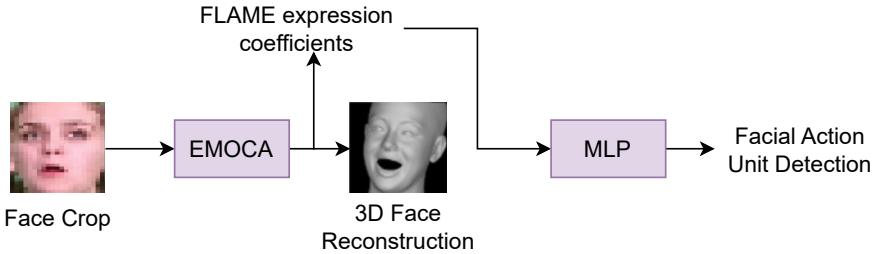


**Figure 6.1** – Overview of the FLAME-based FAU detection pipeline. The input face image is processed by EMOCA to extract FLAME expression coefficients, which are then used by an MLP for multi-label FAU classification.

We train an MLP to perform multi-label classification of FAUs from the Aff-Wild2 dataset, using these FLAME expression coefficients as input. Since multiple FAUs can be activated simultaneously, we formulate this as a multi-label problem and use binary cross-entropy loss as the objective function. We explore MLP architectures with 1 and 2 hidden layers, and select the best-performing configuration based on validation F1 score. The model is optimized using Adam with a learning rate of 0.01, along with learning rate decay for stable convergence. Figure 6.1 illustrates the overview of the FLAME-based FAU detection pipeline. Given the temporal nature of facial movements, we also explore the use of LSTM (Hochreiter et al., 1997) networks to capture temporal dependencies in the FLAME coefficients. To adopt the model to reflect linguistic non-manual features, we finetune the best MLP model obtained by training on Aff-Wild2, on the SMILE-SRT dataset, which includes explicit annotations for non-manual features in sign language, enabling adaptation of the model to linguistically meaningful facial cues.

#### Face-I3D based FAU detection

As a pre-training step, we train an I3D model on the Aff-Wild2 dataset for FAU detection, covering 12 action unit classes. This is formulated as a multi-label classification problem, as multiple FAUs can be activated simultaneously within a single frame. We call this model Face-I3D. The framework is illustrated in Figure 6.2


**Figure 6.2** – Overview of the I3D method for facial action unit detection. The model processes sequences of 16 consecutive cropped face frames, allowing it to capture temporal dynamics in facial expressions.

For face detection, we use MTCNN (Xiang et al., 2017) to locate and crop face regions from individual frames in the Aff-Wild2 dataset. Each face crop is resized to 224×224 pixels. The model takes as input sequences of 16 consecutive cropped face frames, allowing it to model the temporal evolution of facial expressions. To improve generalization, we apply slight data augmentations during training.

The I3D model is trained using a batch size of 16, with binary cross-entropy loss as the objective function. Optimization is performed using Stochastic Gradient Descent (SGD) with a momentum of 0.9 and an initial learning rate of 0.01. The best model is selected based on the Aff-Wild2 validation set F1 score, and it is used for finetuning for non-manual detection.

While FAUs capture important facial cues, their definitions do not directly align with the non-manual features relevant to sign language. Therefore, it is necessary to fine-tune the pre-trained model on a sign language dataset annotated with non-manual features. We use the non-manual annotated SMILE-SRT dataset for this purpose. Fine-tuning is performed for

10 epochs using the Adam (Kingma et al., 2014) optimizer with a learning rate of 0.001 and a batch size of 16.

### 6.1.3 FMAE-based FAU detection

We use the available FMAE (Ning et al., 2025), a model initially trained in a self-supervised manner using a masked image modeling task to learn robust facial representations. Following this pretraining stage, the model is further trained on a supervised FAU classification task using the BP4D (Zhang et al., 2014) dataset, which provides ground truth annotations for multiple facial action units. This two-stage training process enables the model to benefit from both large-scale unsupervised learning and targeted supervised adaptation.

To adapt the model to sign language-specific non-manual cues, we fine-tune only the classifier head (i.e., the final linear layer) on the SMILE-SRT dataset, which includes annotations for non-manual features relevant to sign language. Finetuning is performed using the AdamW (Loshchilov et al., 2017) optimizer with a learning rate of 0.0001, and the model is trained using binary cross-entropy loss for multi-label classification. Figure 6.3 illustrates the approach for FAU detection.

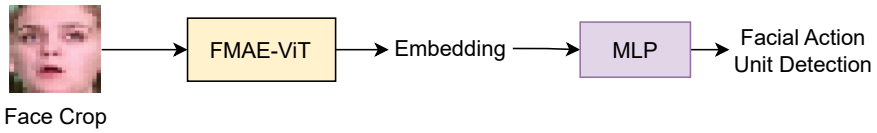


**Figure 6.3** – Overview of the FMAE-based approach. Facial embeddings are extracted using a self-supervised masked autoencoder pretrained and further trained on FAU classification. A linear classifier is fine-tuned on SMILE-SRT for non-manual feature detection.

The annotated L2 signer data from the SMILE-SRT dataset is split into training and validation sets (70:30) and used to fine-tune the non-manual feature detection models.

### 6.1.4 Results

To account for class imbalance, we report the class-weighted F1 score, which adjusts the contribution of each class by weighting its F1 score according to its frequency in the dataset. Figure 6.4 presents the class weighted F1 scores for various models evaluated on the FAU classification task on the Aff-Wild2 dataset. Among the FLAME-based models, the LSTM does slightly better than the MLP variants, demonstrating the benefit of modeling temporal dynamics. However, both Face-I3D and ViT-based FMAE outperform the FLAME-based baselines. Notably, Face-I3D achieves the highest performance with an F1 score of 0.578, followed closely by FMAE at 0.533, indicating the effectiveness of both architectures for capturing expressive facial features. We emphasize that fine-tuning on the SMILE-SRT dataset (which includes sign language-specific non-manual annotations) was conducted only for the Face-I3D and FMAE models, enabling adaptation from general FAU classification to the target

sign language domain.

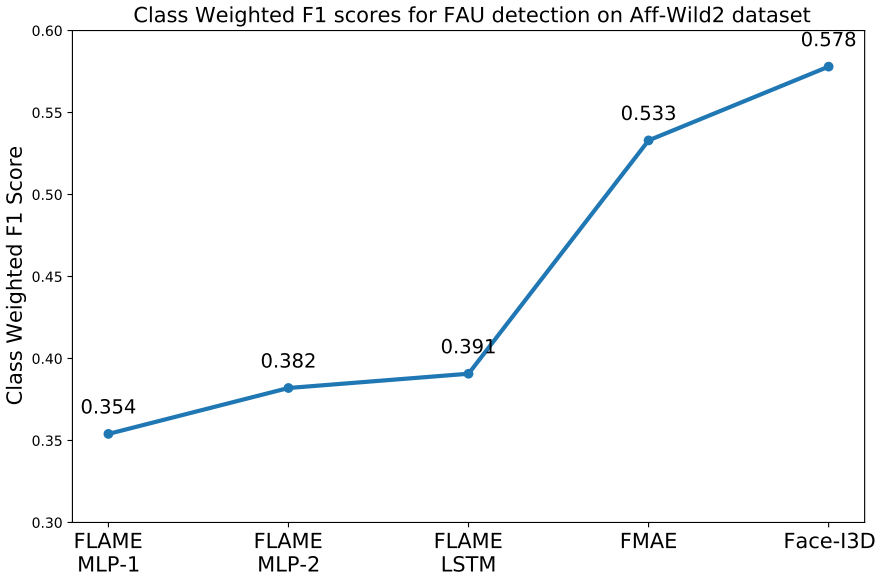


**Figure 6.4** – Performance comparison of FLAME-based, Face-I3D, and FMAE-based FAU detection on the Aff-Wild2 dataset (Class-weighted F1 Score).

Table 6.1 presents the class-weighted F1 scores for Face-I3D and FMAE models after finetuning on the SMILE-SRT dataset, which includes sign language-specific non-manual annotations. Both models demonstrate the ability to adapt from general FAU detection to the domain of sign language.

Face-I3D achieves the best performance with an F1 score of 0.425, slightly outperforming FMAE at 0.407. These results suggest the effectiveness of spatio-temporal facial modeling approaches for non-manual feature detection in sign language videos.

**Table 6.1** – Performance comparison of Face-I3D and FMAE-based NMF detection finetuning on SMILE-SRT dataset (Class-weighted F1 Score). The best performance is highlighted in bold.

| Model | F1 Score |
|---|---|
| Face-I3D | **0.425** |
| FMAE | 0.407 |

In the next section, we integrate Face-I3D and FMAE-based non-manual feature detection into a phonology-based framework for continuous sign language segmentation. We evaluate the impact of these non-manual features on temporal sign segmentation through alignment with and without non-manual cues.

## 6.2 Integration for sign alignment

In moving from isolated to continuous SLA, alignment becomes a critical component. Unlike isolated signs, continuous signing lacks explicit temporal boundaries. Therefore, to exploit isolated SLA methods to continuous settings, we must first infer the start and end of each sign making temporal alignment a prerequisite for lexeme-level assessment. In this context, assessment is framed as a two-step process: first aligning a continuous signing sequence to a series of lexical units (glosses), and then scoring each aligned segment in terms of articulatory quality using manual and non-manual features. In this section, we focus solely on the alignment step.

To validate the contribution of non-manual features to continuous sign language modeling, we analyze their effect on temporal sign segmentation. Specifically, we compare alignment performance under two conditions:

- using only manual features, and
- using a combination of manual and non-manual features.

This evaluation is situated in the context of continuous sign language, where sign boundaries are not explicitly annotated and must be inferred from the data. The validation involves two key steps:

1. Training KL-HMMs on both feature configurations to model the temporal structure of signs.
2. Aligning the trained models to the target sign language data to estimate sign boundaries (start and end times).

We now describe the alignment framework, the features used, and the evaluation metrics.

### 6.2.1 Experimental setup

We adopt a phonology-based framework to model continuous sign sentences from the SMILE-SRT dataset. This approach allows for multi-channel integration of manual and non-manual input channels through their posterior representations, while preserving temporal structure. Each channel contributes complementary linguistic information.

#### Manual features

We investigate two types of posterior representations for hand movement:

- **I3D-based posteriors**: Extracted using a pretrained I3D model and converted to posteriors via a subunit classifier trained on the SMILE-DSGS dataset (isolated signs), as described in Chapter 4.

- **Skeleton-based posteriors**: Derived from MediaPipe hand skeleton features and classified using a pretrained subunit classifier trained on SMILE-DSGS dataset, as described in Chapter 3.

For handshape posteriors, we use MediaPipe hand landmark-based features classified using the subunit method described in Section 5.2.2. The subunit classifier is trained on a curated set of handshape categories for DSGS.

#### Non-manual features

For non-manual features, we explore two distinct approaches for generating posteriors:

- FMAE-based posteriors: We use an FMAE model pretrained and fine-tuned for FAU detection.

- Face-I3D-based posteriors: We extract spatio-temporal features from the face region using an I3D network.

Since both models are trained for multi-label classification, each output dimension corresponds to the likelihood of a particular non-manual feature (e.g., cheeks, eyebrow raise) being present in a frame. Each of these dimensions serve as independent channels in the KL-HMM system. To account for temporal variation across signs, we experiment with models having different numbers of HMM states (e.g., 8 and 20). For each system, we select the best-performing configuration based on validation alignment accuracy and report its results in our final analysis. We use only the L1 signer data from SMILE-SRT for this analysis.

Figure 6.5 presents an overview of our full integration pipeline. It shows how skeleton-based and I3D-based classifiers are used to generate manual feature posteriors, while various non-manual detection strategies produce non-manual posteriors. All channels are combined in a KL-HMM system trained on SMILE-SRT dataset.

### 6.2.2 Evaluation metrics

To evaluate segmentation quality, we use the Jaccard Similarity Score (JSS), which quantifies the overlap between predicted and ground truth sign boundaries. JSS is defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|},$$

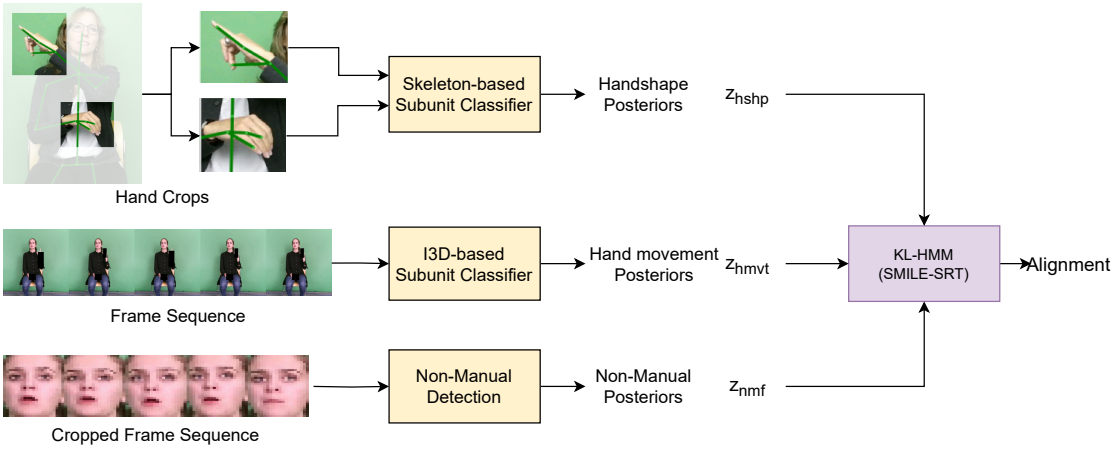


**Figure 6.5** – Overview of the complete system integrating manual and non-manual features for KL-HMM-based alignment. Handshape and hand movement posteriors are extracted using skeleton- and I3D-based subunit classifiers respectively, while facial expression posteriors are obtained via non-manual detection. All channels contribute to a multi-channel KL-HMM for modeling followed by sign segmentation through alignment.

where, $A$ and $B$ are the sets of predicted and ground truth boundaries, respectively, and $0 \leq J(A,B) \leq 1$. A score of 1 indicates perfect alignment; a score of 0 indicates no overlap. JSS is robust to repeated elements and has been widely used in various tasks including keyword similarity (Niwattanakul et al., 2013), recommender systems (Bag et al., 2019), activity recognition (Ivanova et al., 2021), and phonological proximity in sign language (Naranjo-Zeledón et al., 2020). With the manual and non-manual posteriors extracted and integrated into the phonology-based framework, we proceed to evaluate their impact on temporal sign segmentation. These models are trained exclusively on L1 signer data[1], consisting of approximately 12 annotated sentences per signer across 9 signers, meaning that each sentence occurs roughly 9 times in total. While this constitutes a relatively small dataset, it serves as a valuable benchmark to test the robustness of our framework. The next section presents the results across different feature configurations, highlighting the contribution of non-manual cues to segmentation performance.

### 6.2.3 Results and analysis

Table 6.2 reports the Jaccard Similarity Scores for various combinations of manual and non-manual features used within the alignment framework. The following observations can be made from the table:

- When using manual features alone, the I3D-based hand movement representation performs slightly better than MediaPipe-based features, achieving a JSS of 0.568 compared to 0.554, indicating the benefit of spatio-temporal modeling for hand movement.

[1] L1 signers are native signers who acquired sign language as their first language

**Table 6.2** – Jaccard Similarity Scores for sign segmentation for different manual and non-manual feature combinations. In all the cases, the MediaPipe skeleton-based handshape posteriors are used as part of the manual feature stack. Best performance is highlighted in bold.

| Manual Feature | Non-Manual Feature | JSS |
|---|---|---|
| MediaPipe | None | 0.554 |
| MediaPipe | Face-I3D | 0.555 |
| MediaPipe | FMAE | 0.560 |
| I3D | None | 0.568 |
| I3D | Face-I3D | **0.608** |
| I3D | FMAE | 0.556 |

- The integration of non-manual features further improves segmentation performance. Notably, combining I3D manual features with Face-I3D-based non-manual features yields the highest JSS of 0.608, highlighting the complementary role of facial expressions in enhancing alignment accuracy. There no notable performance improvement is observed when adding non-manual information to MediaPipe-based manual features (0.560 vs. 0.555 vs. 0.554).

## 6.3 Summary

In this chapter, we investigated the role of non-manual features particularly facial action units in the context of continuous sign language modeling and assessment. We explored several methods for non-manual feature detection using FAUs as a proxy, and evaluated their effectiveness in both detection and alignment tasks. Experimental results on the Aff-Wild2 dataset suggest that spatio-temporal models (Face-I3D) outperform frame-based FMAE, and FLAME-based MLP and LSTM baselines in FAU detection.

To validate the relevance of non-manual features in sign language segmentation, we integrated the NMF posteriors into a phonology-based alignment framework. Using the SMILE-SRT dataset, we compared models trained with manual features alone to those enhanced with non-manual channels. The results showed consistent improvement in alignment quality when non-manual features especially those from the Face-I3D model.

# 7 Conclusions and future directions

In this thesis, we advanced phonology-based sign language assessment in two key contexts: to support learner feedback and to evaluate the quality of machine-generated signing.

In Chapter 3, we evaluated 2D and 3D skeleton estimation methods derived from RGB input with the objective of enabling accessible, webcam-based sign language learning and assessment. The primary goal was to assess the impact of losing depth information present in systems like Kinect on both recognition and assessment performance. Our findings show that while RGB-based skeletons provide an interpretable and modular representation of signing, their effectiveness is constrained by the accuracy and consistency of pose estimation models. Importantly, we observed that recognition accuracy and assessment quality are not always correlated. A method achieving high recognition performance may still fail to capture articulatory quality in a way that supports reliable assessment and vice-versa. This highlights the necessity of evaluating recognition and assessment performance separately when designing sign language learning tools. Despite the absence of depth data, RGB-based methods achieved competitive assessment results, suggesting that depth is not strictly required for effective evaluation of signing performance. The methods developed in this chapter were integrated into a web-based sign language learning application, which was presented as a demo at an international conference (Tornay et al., 2023). Subsequent work built on this system to conduct rater studies (Holzknecht et al., 2024) comparing human judgments and automatic assessment outputs, further validated its practical relevance. Furthermore, the RGB-based assessment system developed is being integrated to support the Swiss Deaf Association as part of Inclusive Information and Communication Technologies (IICT) project.

In Chapter 4, we investigated the use of deep learning-based representations for sign language assessment, with a particular focus on evaluating hand movement and handshape quality in isolated signing. Two categories of models were examined: supervised convolutional networks (I3D) and self-supervised Vision Transformers (ViTs). We analyzed their effectiveness in capturing articulatory dynamics relevant to assessment. Our findings show that I3D-based features provide a strong alternative to skeleton-based representations, particularly in modeling fine-grained movement patterns. In contrast, ViT-based models although effective for

sign recognition demonstrated noticeably weaker performance in assessment tasks. This gap reinforces the observation from Chapter 3: high recognition accuracy does not necessarily imply strong assessment capability. This work also underscores the trade-offs between supervised, task-specific models and sign language agnostic, pre-trained representations. While sign language agnostic features may underperform in detailed assessment, they still encode meaningful motion information and may serve as practical alternatives in low-resource or cross-lingual scenarios. Overall, the results establish the viability of deep spatio-temporal features as a foundation for phonologically grounded sign language assessment. Building on these insights, we then investigate how such feature representations (video-based and pose-based) can be used to assess automatically generated sign language content.

In Chapter 5, as a step toward integrating feedback into sign language generation systems in a manner aligned with human judgment, we proposed a posterior-based evaluation framework that leverages phonological posteriors specifically for hand movement and handshape to assess the articulatory quality of generated signing outputs. By applying this framework to both video-to-video and text-to-pose generation models, we conducted a detailed comparison between generated and reference sign productions. Importantly, the framework is modality-flexible, it can be applied to evaluate both RGB video and pose-based outputs using the methods developed in Chapter 3 and Chapter 4, making it suitable for a wide range of generative architectures. This approach draws inspiration from prior work in speech assessment, where posterior-based methods have been successfully used to evaluate speech intelligibility in pathological speech (Fritsch et al., 2021) and accentedness detection in non-native speakers (Rasipuram et al., 2015). These approaches, originally developed in the context of spoken language processing, have been successfully extended to sign language in this thesis. This cross-modal transfer highlights the broader applicability of phonologically grounded, posterior-based assessment frameworks, which align well with human judgment, to human language technologies across modalities. Notably, the models used for posterior extraction were trained on a different sign language than the one used in the assessment evaluation. This cross-lingual setup suggests that the posterior components may exhibit a degree of language independence, indicating the potential of the framework to generalize across sign languages, an important step toward building robust, scalable assessment tools. Together, these results support the use of posterior-based evaluation as a linguistically meaningful and interpretable alternative for assessing generated outputs. They also highlight the potential of such metrics to serve not only as diagnostic tools but also as feedback mechanisms within future generation pipelines, helping to improve the quality and usability of automatically generated sign language content.

Finally, in Chapter 6, we took an initial step towards the transition from isolated to continuous sign language assessment. Until this point, our assessment framework focused exclusively on manual features. However, in continuous signing, non-manual cues such as facial expressions and head movements play a crucial role in signaling grammatical structure and marking sign boundaries. Despite their linguistic importance, non-manual feature detection remains an under-researched problem in the field. In this context, we developed and evaluated methods

for non-manual feature detection by training models on the large-scale Aff-Wild2 FAU dataset and fine-tuning them on the SMILE-SRT sign language dataset. Among the models explored, those using spatio-temporal representations yielded the best performance in detecting non-manual cues relevant to signing. We then successfully integrated the resulting non-manual posteriors into our phonology-based framework for continuous sign segmentation. The framework incorporates both manual and non-manual components. The manual posteriors are derived from skeleton-based and spatio-temporal features, as described in Chapters 3 and 4. Importantly, although the posterior models were trained on isolated sign language data, they generalized effectively to continuous signing, enabling segmentation without prior alignment. Through a comparative analysis of alignment results with and without non-manual channels we demonstrated that incorporating non-manual features improves segmentation accuracy. These findings underscore the importance of non-manual cues in effective identification of sign boundaries.

We suggest the following future directions:

- **Sign language aware transformer-based assessment**: While ViT-based features have demonstrated promising results in recognition tasks, their effectiveness in articulatory assessment remains limited compared to sign language aware systems. A promising direction is to adapt transformer-based models for sign language assessment by incorporating sign language aware posteriors derived from domain-specific subunits, such as handshape or movement categories. This approach could enhance the model's ability to capture the nuances of sign language articulation, potentially bridging the performance gap observed in our studies.

- **Continuous sign language assessment**: The current methodology lays the groundwork for continuous sign language assessment by first performing sign segmentation, with the goal of subsequently applying isolated sign assessment techniques to each segmented unit. In this thesis, we focused specifically on the segmentation component, demonstrating that phonology-based models augmented with non-manual features can improve alignment quality in continuous signing, but the assessment of the segmented signs is still an open problem. A key future direction is to extend this framework by using the segmented sign boundaries to exploit isolated sign language assessment within continuous signing sequences. Another open research problem is to move toward global continuous sign language assessment, that is, assessing signing performance directly on continuous sequences without relying on prior segmentation.

- **Fluency assessment**: While this thesis focused on phonological assessment in a reference-based setting, an important future direction is the evaluation of fluency: how naturally and confidently a signer produces sequences of signs. Fluency plays a crucial role in overall proficiency and communicative effectiveness, particularly in free or spontaneous signing, where no explicit reference is available. Extending the phonology-based framework to assess fluency would require systems that jointly perform recognition

and assessment, enabling the analysis of temporal smoothness, co-articulation, and sign transitions. This shift raises new challenges, especially in managing the trade-off between recognition accuracy and assessment reliability mentioned earlier.

# Bibliography


Adaloglou, Nikolas et al. (2021). “A comprehensive study on deep learning-based methods for sign language recognition”. In: *IEEE Transactions on Multimedia* 24, pp. 1750–1762 (cit. on p. 8).

Aghabozorgi, Saeed, Ali Seyed Shirkhorshidi, and Teh Ying Wah (2015). “Time-Series Clustering–A Decade Review”. In: *Information systems* 53, pp. 16–38 (cit. on p. 59).

Albanie, Samuel et al. (2020). “BSL-1K: Scaling Up Co-articulated Sign Language Recognition Using Mouthing Cues”. In: *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*. Glasgow, United Kingdom: Springer-Verlag, pp. 35–53 (cit. on p. 66).

Aradilla, Guillermo, Herve Bourlard, and Mathew Magimai.-Doss (2008). “Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task”. In: *Proceedings of Interspeech*, pp. 928–931 (cit. on p. 12).

Aradilla, Guillermo, Jithendra Vepa, and Herve Bourlard (2007). “An Acoustic Model Based on Kullback-Leibler Divergence for Posterior Features”. In: *ICASSP*, pp. 657–660 (cit. on p. 12).

Aran, Oya et al. (2009). “SignTutor: An Interactive System for Sign Language Tutoring”. In: *IEEE MultiMedia* 16.1, pp. 81–93 (cit. on p. 9).

Arendsen, Jeroen et al. (2008). “Acceptability Ratings by Humans and Automatic Gesture Recognition for Variations in Sign Productions”. In: *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 1–6 (cit. on pp. 30, 33).

Arkushin, Rotem Shalev, Amit Moryossef, and Ohad Fried (2023). “Ham2Pose: Animating Sign Language Notation into Pose Sequences”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21046–21056 (cit. on pp. xviii, 9, 54, 58, 60, 62).

Arnab, Anurag, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid (2021). “ViViT: A Video Vision Transformer”. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 6816–6826 (cit. on p. 33).

Asteriadis, Stylianos, George Caridakis, and Kostas Karpouzis (2012). “Non-Manual Cues in Automatic Sign Language Recognition”. In: *Personal and Ubiquitous Computing* 18 (cit. on p. 66).

Bag, Sujoy, Sri Krishna Kumar, and Manoj Kumar Tiwari (2019). “An Efficient Recommendation Generation Using Relevant Jaccard Similarity”. In: *Information Sciences* 483, pp. 53–64 (cit. on p. 74).

Baltatzis, Vasileios, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou (2024). “Neural Sign Actors: A Diffusion Model for 3D Sign Language Production from Text”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1985–1995 (cit. on p. 9).

Bazarevsky, Valentin, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann (2020). “BlazePose: On-device Real-time Body Pose tracking”. In: *ArXiv* abs/2006.10204 (cit. on p. 23).

Blattmann, Andreas et al. (2023). *Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets* (cit. on p. 2).

Bradski, Gary (2000). “The OpenCV Library”. In: *Dr. Dobb's Journal of Software Tools* (cit. on p. 54).

Brashear, Helene et al. (2006). “American Sign Language Recognition in Game Development for Deaf Children”. In: *Proc. of the International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 79–86 (cit. on p. 9).

Bull, Hannah, Triantafyllos Afouras, Gul Varol, Samuel Albanie, Liliane Momeni, and Andrew Zisserman (2021). “ Aligning Subtitles in Sign Language Videos ”. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 11532–11541 (cit. on p. 9).

Camgoz, Necati Cihan, Simon Hadfield, Oscar Koller, and Richard Bowden (2017). “SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3075–3084 (cit. on pp. 8, 25).

Camgoz, Necati Cihan, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden (2018). “Neural Sign Language translation”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7784–7793 (cit. on p. 8).

Camgoz, Necati Cihan, Oscar Koller, Simon Hadfield, and Richard Bowden (2020). “Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on p. 8).

Cao, Zhe, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh (2019). “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (cit. on pp. 21, 23, 54).

Carreira, João and Andrew Zisserman (2017). “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733 (cit. on pp. 33, 34).

Chan, Caroline, Shiry Ginosar, Tinghui Zhou, and Alexei Efros (2019). “Everybody Dance Now”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5932–5941 (cit. on p. 52).

Cooper, Helen, Brian Holt, and Richard Bowden (2011). “Sign Language Recognition”. In: *Visual Analysis of Humans, 2011* (cit. on p. 8).

Cooper, Helen, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden (2012). "Sign Language Recognition using Sub-Units". In: *Journal of Machine Learning Research* 13, pp. 2205–2231 (cit. on p. 65).

Corneanu, Ciprian, Meysam Madadi, and Sergio Escalera (2018). "Deep Structure Inference Network for Facial Action Unit Recognition". In: *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XII*. Munich, Germany: Springer-Verlag, pp. 309–324 (cit. on p. 68).

Cory, Oliver et al. (2024). *Modelling the Distribution of Human Motion for Sign Language Assessment* (cit. on pp. 10, 18, 33, 59).

Dan, Jun et al. (2023). "TransFace: Calibrating Transformer Training for Face Recognition from a Data-Centric Perspective". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20585–20596 (cit. on p. 68).

Danecek, Radek, Michael J. Black, and Timo Bolkart (2022). "EMOCA: Emotion Driven Monocular Face Capture and Animation". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20311–20322 (cit. on p. 67).

De Coster, Mathieu, Mieke Van Herreweghe, and Joni Dambre (2020). "Sign Language Recognition with Transformer Networks". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 6018–6024 (cit. on p. 8).

Dosovitskiy, Alexey et al. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations* (cit. on p. 35).

Ebling, Sarah et al. (2018). "SMILE Swiss German Sign Language Dataset". In: *Proc. of the Language Resources and Evaluation Conference* (cit. on p. 16).

Ekman, Paul and Wallace V. Friesen (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press (cit. on p. 66).

Feng, Yao, Haiwen Feng, Michael J. Black, and Timo Bolkart (2021). "Learning an Animatable Detailed 3D Face Model from In-The-Wild Images". In: *ACM Transactions on Graphics (ToG), Proc. SIGGRAPH* 40.8 (cit. on p. 67).

Fleiss, Joseph L. (1971). "Measuring Nominal Scale Agreement Among Many Raters". In: *Psychological bulletin* 76.5, p. 378 (cit. on p. 61).

Freitag, Markus et al. (2022). "Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust". In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Ed. by Philipp Koehn et al. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 46–68 (cit. on p. 50).

Fritsch, Julian and Mathew Magimai-Doss (2021). "Utterance Verification-Based Dysarthric Speech Intelligibility Assessment Using Phonetic Posterior Features". In: *IEEE Signal Processing Letters* 28, pp. 224–228 (cit. on p. 78).

Goodfellow, Ian et al. (2020). "Generative Adversarial Networks". In: *Commun. ACM* 63.11, pp. 139–144 (cit. on p. 2).

He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick (2017). "Mask R-CNN". In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988 (cit. on pp. 21, 23).

Hessel, Jack, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi (2021). "CLIPScore: A Reference-free Evaluation Metric for Image Captioning". In: *EMNLP* (cit. on p. 60).

Hieber, Felix et al. (2022). "Sockeye 3: Fast Neural Machine Translation with Pytorch". In: *arXiv preprint arXiv:2207.05851* (cit. on p. 58).

Higgins, Irina et al. (2017). "beta-vae: Learning basic visual concepts with a constrained variational framework". In: *International conference on learning representations* (cit. on p. 59).

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Comput.* 9.8, pp. 1735–1780 (cit. on p. 69).

Holzknecht, Franz et al. (2024). "Automated Sign Language Vocabulary Assessment: Comparing Human and Machine Ratings and Studying Learner Perceptions". In: *Language Assessment Quarterly* 21.3. Publisher Copyright: © 2024 The Author(s). Published with license by Taylor & Francis Group, LLC., pp. 245–265 (cit. on p. 77).

Hu, Hezhen, Wengang Zhou, Junfu Pu, and Houqiang Li (2021). "Global-Local Enhancement Network for NMF-Aware Sign Language Recognition". In: *ACM Trans. Multimedia Comput. Commun. Appl.* 17.3 (cit. on p. 66).

Huenerfauth, Matt, Elaine Gale, Brian Penly, Sree Pillutla, Mackenzie Willard, and Dhananjai Hariharan (2017). "Evaluation of Language Feedback Methods for Student Videos of American Sign Language". In: *ACM Trans. Access. Comput.* 10.1 (cit. on p. 10).

Ionescu, Catalin, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu (2014). "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7, pp. 1325–1339 (cit. on p. 23).

ISARA, Application (2016). *ISARA app*. URL: https://isara.app/features (cit. on p. 10).

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros (2017). "Image-to-Image Translation with Conditional Adversarial Networks". In: *CVPR* (cit. on pp. 53, 54).

Ivanova, Iustina, Marina Andric, Andrea Janes, Francesco Ricci, and Floriano Zini (2021). "Climbing Activity Recognition and Measurement with Sensor Data Analysis". In: *Companion Publication of the 2020 International Conference on Multimodal Interaction*. ICMI '20 Companion (cit. on p. 74).

Ivashechkin, Maksym, Oscar Mendez, and Richard Bowden (2023). "Improving 3D Pose Estimation For Sign Language". In: *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pp. 1–5 (cit. on p. 59).

Jiang, Zifan, Colin Leong, and Amit Moryossef (2024). *Pose Evaluation: Metrics for Evaluating Sign Language Generation Models*. https://github.com/sign-language-processing/pose-evaluation (cit. on p. 50).

Jiang, Zifan, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling (2024). "SignCLIP: Connecting Text and Sign Language by Contrastive Learning". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed.

by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 9171–9193 (cit. on p. 60).

John, Christopher (2012). "SignAssess – Online Sign Language Training Assignments via the Browser, Desktop and Mobile". In: *Computers Helping People with Special Needs*. Ed. by Klaus Miesenberger, Arthur Karshmer, Petr Penaz, and Wolfgang Zagler. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 253–260 (cit. on p. 9).

Kay, Will et al. (2017). "The Kinetics Human Action Video Dataset". In: (cit. on p. 37).

Kingma, Diederik P. and Jimmy Ba (2014). "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980 (cit. on p. 70).

Kocabas, Muhammed, Nikos Athanasiou, and Michael J. Black (2020). "VIBE: Video Inference for Human Body Pose and Shape Estimation". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on pp. 22, 24).

Koller, Oscar, Hermann Ney, and Richard Bowden (2016). "Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled". In: *IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA, pp. 3793–3802 (cit. on p. 26).

Kollias, Dimitrios (2022). "ABAW: Valence-Arousal Estimation, Expression Recognition, Action Unit Detection & Multi-Task Learning Challenges". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2328–2336 (cit. on p. 17).

Kollias, Dimitrios (2023a). "ABAW: Learning from Synthetic Data & Multi-Task Learning Challenges". In: *European Conference on Computer Vision*. Springer, pp. 157–172 (cit. on p. 17).

Kollias, Dimitrios, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou (2020). "Analysing Affective Behavior in the First ABAW 2020 Competition". In: *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pp. 794–800 (cit. on p. 17).

Kollias, Dimitrios, Viktoriia Sharmanska, and Stefanos Zafeiriou (2019). "Face Behavior a la carte: Expressions, Affect and Action Units in a Single Network". In: *arXiv preprint arXiv:1910.11111* (cit. on p. 17).

Kollias, Dimitrios, Viktoriia Sharmanska, and Stefanos Zafeiriou (2021). "Distribution Matching for Heterogeneous Multi-Task Learning: a Large-scale Face Study". In: *arXiv preprint arXiv:2105.03790* (cit. on p. 17).

Kollias, Dimitrios, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou (2023b). "ABAW: Valence-Arousal Estimation, Expression Recognition, Action Unit Detection & Emotional Reaction Intensity Estimation Challenges". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5888–5897 (cit. on pp. 17, 68).

Kollias, Dimitrios, Panagiotis Tzirakis, et al. (2019). "Deep Affect Prediction in-the-wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond". In: *International Journal of Computer Vision*, pp. 1–23 (cit. on p. 17).

Kollias, Dimitrios and Stefanos Zafeiriou (2019). "Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace". In: *arXiv preprint arXiv:1910.04855* (cit. on p. 17).

Kollias, Dimitrios and Stefanos Zafeiriou (2021a). "Affect Analysis in-the-wild: Valence-Arousal, Expressions, Action Units and a Unified Framework". In: *arXiv preprint arXiv:2103.15792* (cit. on p. 17).

Kollias, Dimitrios and Stefanos Zafeiriou (2021b). "Analysing Affective Behavior in the second ABAW2 Competition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3652–3660 (cit. on p. 17).

Konrad, Reiner et al. (2020). *MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release*. Version 3.0 (cit. on pp. 34, 36).

Kratimenos, Agelos, Georgios Pavlakos, and Petros Maragos (2021). "Independent Sign Language Recognition with 3d Body, Hands, and Face Reconstruction". In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4270–4274 (cit. on p. 66).

Krishna, Shyam, Janmesh Ukey, and Dinesh Babu Jayagopi (2021). "GAN-Based Indian Sign Language Synthesis". In: *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*. ICVGIP '21. Jodhpur, India: Association for Computing Machinery (cit. on pp. xiv, 50, 52, 53).

Li, Tianye, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero (2017). "Learning a Model of Facial Shape and Expression From 4D Scans". In: *ACM Trans. Graph.* 36.6 (cit. on p. 67).

Li, Yanghao et al. (2022). "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection". In: *CVPR* (cit. on p. 35).

Liu, Shuying and Weihong Deng (2015). "Very Deep Convolutional Neural Network Based Image Classification Using Small Training Sample Size". In: *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 730–734 (cit. on p. 54).

Liu, Ze et al. (2021). "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (cit. on pp. 35, 68).

Loper, Matthew, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black (2015). "SMPL: A Skinned Multi-Person Linear Model". In: *ACM Trans. Graph.* 34.6 (cit. on p. 24).

Loshchilov, Ilya and Frank Hutter (2017). "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations* (cit. on p. 70).

Lugaresi, Camillo et al. (2019). "MediaPipe: A Framework for Perceiving and Processing Reality". In: *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019* (cit. on pp. 22, 54).

Ma, Fuyan, Bin Sun, and Shutao Li (2023). "Facial Expression Recognition With Visual Transformers and Attentional Selective Fusion". In: *IEEE Transactions on Affective Computing* 14.2, pp. 1236–1248 (cit. on p. 68).

Mercanoglu, Ozge and Hacer Keles (2022). "Using Motion History Images With 3D Convolutional Networks in Isolated Sign Language Recognition". In: *IEEE Access* 10, pp. 1–1 (cit. on p. 8).

Moryossef, Amit (2024). “sign.mt: Real-Time Multilingual Sign Language Translation Application”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Delia Irazu Hernandez Farias, Tom Hope, and Manling Li. Miami, Florida, USA: Association for Computational Linguistics, pp. 182–186 (cit. on p. 9).

Moryossef, Amit, Mathias Müller, Anne Göhring, Zifan Jiang, Yoav Goldberg, and Sarah Ebling (2023). “An Open-Source Gloss-Based Baseline for Spoken to Signed Language Translation”. In: *Proceedings of the Second International Workshop on Automatic Translation for Signed and Spoken Languages*. Ed. by Dimitar Shterionov et al. Tampere, Finland: European Association for Machine Translation, pp. 22–33 (cit. on p. 58).

Moryossef, Amit, Kayo Yin, Graham Neubig, and Yoav Goldberg (2021). “Data Augmentation for Sign Language Gloss Translation”. In: *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*. Ed. by Dimitar Shterionov. Virtual: Association for Machine Translation in the Americas, pp. 1–11 (cit. on p. 8).

Mukushev, Medet, Arman Sabyrov, Alfarabi Imashev, Kenessary Koishybay, Vadim Kimmelman, and Anara Sandygulova (2020). “Evaluation of Manual and Non-manual Components for Sign Language Recognition”. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 6073–6078 (cit. on p. 65).

Müller, Mathias, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling (2023). “Considerations for Meaningful Sign Language Machine Translation Based on Glosses”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 682–693 (cit. on p. 8).

Naranjo-Zeledón, Luis, Mario Chacón-Rivas, Jesús Peral, and Antonio Ferrández (2020). “Phonological Proximity in Costa Rican Sign Language”. In: *Electronics* (cit. on p. 74).

Neves, Carolina, Luísa Coheur, and Hugo Nicolau (2020). “HamNoSyS2SiGML: Translating HamNoSys into SiGML”. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 6035–6039 (cit. on p. 9).

Ning, Mang, Albert Ali Salah, and Itir Onal Ertugrul (2025). *Representation Learning and Identity Adversarial Training for Facial Behavior Understanding* (cit. on pp. 68, 70).

Niwattanakul, Suphakit, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu (2013). “Using the Jaccard Coefficient for Keywords Similarity”. In: *Proceedings of the international multiconference of engineers and computer scientists* (cit. on p. 74).

Oquab, Maxime et al. (2023). *DINOv2: Learning Robust Visual Features without Supervision* (cit. on pp. 34, 35, 37, 68).

Papadimitriou, Katerina and Gerasimos Potamianos (2023a). “Sign Language Recognition via Deformable 3D Convolutions and Modulated Graph Convolutional Networks”. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (cit. on p. 8).

Papadimitriou, Katerina, Gerasimos Potamianos, et al. (2023b). "Greek sign language recognition for an education platform". In: *Univers. Access Inf. Soc.* 24.1, pp. 51–68 (cit. on p. 10).

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "BLEU: A Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Ed. by Pierre Isabelle, Eugene Charniak, and Dekang Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318 (cit. on p. 50).

Pavllo, Dario, Christoph Feichtenhofer, David Grangier, and Michael Auli (2019). "3D Human Pose Estimation in Video with Temporal Convolutions and Semi-Supervised Training". In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on pp. 22, 23).

Pyumina, Elena, Denis Dresvyanskiy, and Alexey Karpov (2022). "In Search of a Robust Facial Expressions Recognition Model: A Large-Scale Visual Cross-Corpus Study". In: *Neurocomputing* 514 (cit. on p. 68).

Rasipuram, Ramya, Milos Cernak, Alexandre Nanchen, and Mathew Magimai-Doss (2015). "Automatic Accentedness Evaluation of Non-Native Speech Using Phonetic and Sub-Phonetic Posterior Probabilities". In: *16th Annual Conference of the International Speech Communication Association, INTERSPEECH 2015, Dresden, Germany, September 6-10, 2015*. ISCA, pp. 648–652 (cit. on p. 78).

Rasipuram, Ramya and Mathew Magimai.-Doss (2016). "Articulatory Feature Based Continuous Speech Recognition Using Probabilistic Lexical Modeling". In: *Comput. Speech Lang.* 36.C, pp. 233–259 (cit. on p. 12).

Rothkrantz, Leon, Dragos Datcu, and Pascal Wiggers (2009). "FACS-Coding of Facial Expressions". In: *Proceedings of the International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing*. CompSysTech '09. Ruse, Bulgaria: Association for Computing Machinery (cit. on p. 66).

Ryali, Chaitanya et al. (2023). "Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles". In: *ICML* (cit. on pp. 33–35, 68).

Sarhan, Noha and Simone Frintrop (2020). "Transfer Learning For Videos: From Action Recognition To Sign Language Recognition". In: *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 1811–1815 (cit. on p. 8).

Sarhan, Noha and Simone Frintrop (2023). "Unraveling a Decade: A Comprehensive Survey on Isolated Sign Language Recognition". In: *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3202–3211 (cit. on p. 8).

Shrout, Patrick E and Joseph L Fleiss (1979). "Intraclass Correlations : Uses in Assessing Rater Reliability". In: *Psychological bulletin* 86.2, p. 420 (cit. on p. 56).

SignAll Technologies Inc. (2021). *SignAll*. https://www.signall.us/. Accessed: 2021-02-27 (cit. on p. 10).

Silva, Emely Pujólli da, Paula Dornhofer Paro Costa, Kate Mamhy Oliveira Kumada, and José Mario De Martino (2022). "Facial Action Unit Detection Methodology With Application in Brazilian Sign Language Recognition". In: *Pattern Anal. Appl.* 25.3, pp. 549–565 (cit. on p. 66).

Spaai, Gerard. et al. (2005). “Elo: An Electronic Learning Environment for Practising Sign Vocabulary by Young Deaf Children”. In: *Proc. of International Congress for Education of the Deaf* (cit. on p. 9).

Stoll, Stephanie, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden (2020). “Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks”. In: *Int. J. Comput. Vision* 128.4, pp. 891–908 (cit. on pp. 9, 50).

Stoll, Stephanie, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden (2018). “Sign Language Production using Neural Machine Translation and Generative Adversarial Networks”. In: *29th British Machine Vision Conference (BMVC 2018)*. Northumbria University, Newcastle Upon Tyne, UK: British Machine Vision Association (cit. on pp. 9, 50).

Sutskever, Ilya, James Martens, George Dahl, and Geoffrey Hinton (2013). “On the Importance of Initialization and Momentum in Deep Learning”. In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. ICML’13. Atlanta, GA, USA: JMLR.org, III–1139–III–1147 (cit. on p. 36).

Tarigopula, Neha, Preyas Garg, Skanda Muralidhar, Sandrine Tornay, Dinesh Babu Jayagopi, and Mathew Magimai.-Doss (2024). “Content-Based Objective Evaluation of Artificially Generated Sign Language Videos”. In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3815–3819 (cit. on p. 50).

Tarigopula, Neha, Sandrine Tornay, Skanda Muralidhar, and Mathew Magimai Doss (2022). “Towards Accessible Sign Language Assessment and Learning”. In: *Proceedings of the 2022 International Conference on Multimodal Interaction*. ICMI ’22. Bengaluru, India: Association for Computing Machinery, pp. 626–631 (cit. on pp. 22, 33, 43, 44, 46).

Tarigopula, Neha, Sandrine Tornay, Ozge Mercanoglu Sincan, Richard Bowden, and Mathew Magimai.-Doss (2025). “Posterior-Based Analysis of Spatio-Temporal Features for Sign Language Assessment”. In: *IEEE Open Journal of Signal Processing* 6, pp. 284–292 (cit. on pp. 34, 36).

Tellamekala, Mani Kumar, Ömer Sümer, Björn W. Schuller, Elisabeth André, Timo Giesbrecht, and Michel Valstar (2024). “Are 3D Face Shapes Expressive Enough for Recognising Continuous Emotions and Action Unit Intensities?” In: *IEEE Transactions on Affective Computing* 15.2, pp. 535–548 (cit. on p. 68).

Tornay, Sandrine (2021). “Explainable Phonology-based Approach for Sign Language Recognition and Assessment”. PhD thesis. Lausanne: IEL, p. 156 (cit. on pp. 10, 13, 51).

Tornay, Sandrine, Oya Aran, and Mathew Magimai.-Doss (2020). “An HMM Approach with Inherent Model Selection for Sign Language and Gesture Recognition”. In: *Proc. of the International Conference on Language Resources and Evaluation LREC 2020* (cit. on p. 39).

Tornay, Sandrine, Necati Cihan Camgoz, Richard Bowden, and Mathew Magimai.-Doss (2020). “A Phonology-based Approach for Isolated Sign Production Assessment in Sign Language”. In: *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI ’20 Companion)* (cit. on pp. 10, 13, 15, 21, 33, 39, 43, 44, 46).

Tornay, Sandrine and Mathew Magimai-Doss (2025). “Towards Dynamic Skeleton-based Handshape Subunits for Sign Language Assessment”. In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (cit. on p. 55).

Tornay, Sandrine and Mathew Magimai.-Doss (2019). “Subunits Inference and Lexicon Development Based on Pairwise Comparison of Utterances and Signs”. In: *Information* 10 (cit. on p. 24).

Tornay, Sandrine, Marzieh Razavi, Necati Cihan Camgoz, Richard Bowden, and Mathew Magimai.-Doss (2019). “HMM-based Approaches to Model Multichannel Information in Sign Language Inspired from Articulatory Features-based Speech Processing”. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2817–2821 (cit. on p. 12).

Tornay, Sandrine et al. (2023). “Web SMILE Demo: A Web Application Providing Automated Feedback on Sign Language Vocabulary Production”. In: *44th Language Testing and Research Colloquium: Language Assessment for a Global, Digital, and More Equitable Era*. Demo presentation. New York City, USA (cit. on p. 77).

Ullmann, Raphael, Mathew Magimai-Doss, and Hervé Bourlard (2015). “Objective Speech Intelligibility Assessment Through Comparison of Phoneme Class Conditional Probability Sequences”. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4924–4928 (cit. on p. 51).

Vasani, Neel, Pratik Autee, Samip Kalyani, and Ruhina Karani (2020). “Generation of Indian Sign Language by Sentence Processing and Generative Adversarial Networks”. In: *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pp. 1250–1255 (cit. on p. 50).

Vaswani, Ashish et al. (2017). “Attention Is All You Need”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., pp. 6000–6010 (cit. on p. 35).

Viegas, Carla, Mert Inan, Lorna Quandt, and Malihe Alikhani (2023). “Including Facial Expressions in Contextual Embeddings for Sign Language Generation”. In: *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*. Ed. by Alexis Palmer and Jose Camacho-collados. Toronto, Canada: Association for Computational Linguistics, pp. 1–10 (cit. on p. 66).

Vogler, Christian and Dimitris Metaxas (1998). “ASL Recognition Based on a Coupling Between HMMs and 3D Motion Analysis”. In: IEEE Computer Society, pp. 363–369 (cit. on p. 8).

Vogler, Christian and Dimitris Metaxas (1999). “Parallel Hidden Markov Models for American Sign Language Recognition”. In: *Proc. of the Seventh IEEE International Conference on Computer Vision (ICCV)*. Vol. 1, 116–122 vol.1 (cit. on p. 8).

Willoughby, Louisa, Stephanie Linder, Kirsten Ellis, and Julie Fisher (2015). “Errors and Feedback in the Beginner Auslan Classroom”. In: *Sign Language Studies* 15, pp. 322–347 (cit. on p. 10).

Xiang, Jia and Gengming Zhu (2017). “Joint Face Detection and Facial Expression Recognition with MTCNN”. In: *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, pp. 424–427 (cit. on pp. 67–69).

Zafeiriou, Stefanos, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia (2017). "Aff-Wild: Valence and Arousal 'In-the-Wild' Challenge". In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, pp. 1980–1987 (cit. on p. 17).

Zafrulla, Zahoor et al. (2011). "CopyCat: An American Sign Language Game for Deaf Children". In: *IEEE International Conference on Automatic Face and Gesture Recognition* (cit. on p. 9).

Zhang, Biao, Garrett Tanzer, and Orhan Firat (2024). "Scaling Sign Language Translation". In: *arXiv preprint arXiv:2407.11855* (cit. on p. 61).

Zhang, Xing et al. (2014). "BP4D-Spontaneous: A High-Resolution Spontaneous 3D Dynamic Facial Expression Database". In: *Image and Vision Computing* 32.10. Best of Automatic Face and Gesture Recognition 2013, pp. 692–706 (cit. on p. 70).

Zhong, Yaoyao and Weihong Deng (2021). *Face Transformer for Recognition* (cit. on p. 68).

Zhou, Benjia et al. (2023). "Gloss-free Sign Language Translation: Improving from Visual-Language Pretraining". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20814–20824. DOI: 10.1109/ICCV51070.2023.01908 (cit. on p. 8).

# NEHA TARIGOPULA

**Ph.D. Candidate**

Martigny, Switzerland | Google Scholar | neha.tarigopula@idiap.ch | nehatarigopula

## EXPERIENCE

### Ph.D. Student in Sign Language Recognition and Assessment

**Idiap & EPFL**

Mar 2021 – Ongoing | Martigny, Switzerland

- PhD Thesis: Towards Continuous Sign Language Assessment and Learning. Supervised by Dr. Mathew Magimai-Doss and Prof. Jean-Marc Odobez.
  The research aims to bridge the communication gap between individuals with hearing disabilities and others by developing a platform for sign language learning by offering detailed feedback on handshape, movement, and facial expressions to support learners.
- Topics: Video content analysis, sign language recognition, video generation, content-based analysis of videos, hand pose classification, keyword spotting, facial expression analysis.

### Research Assistant

**Indian Institute of Technology, Hyderabad**

Jun 2020 – Jan 2021 | Hyderabad, India

Supervisor: Prof. Vineeth Balasubramanian

- Worked on **self-supervised** learning with a focus on image classification and explored unsupervised domain generalization through self-supervision.

### Machine Learning Researcher

**Mercedes Benz Research & Development**

Aug 2017 – May 2020 | Bangalore, India

*Worked on Human Centered Machine Learning that focussed on 2D pose estimation and pose classification to facilitate Smart Interiors in cars*

- Developed a **ResNet**-based hand pose classification model that was deployed in an edge device for favorite function selection in 2021 Mercedes EQS.
- Filed a patent: *Tarigopula, N. (2019, March). Hand pose classification using CNN on depth-based images with multiple hands. (Patent 2019P01559)*
- Led the module for pose classification, overseeing the entire process from defining data and annotation requirements, to model development and quantization, and testing in real-world scenarios.

### Computer Vision Intern

**National Instruments R&D**

May 2015 – Jul 2015 | Bangalore, India

*Hardware optimization of image processing algorithms*

- Optimized polynomial-based flat-field correction algorithm to make the best use of pipe-lining and parallel processing that are fully supported by FPGAs. Achieved 30% speed up in the execution time.

## SKILLS

Python | Pytorch | W&B | Git | Lightning | HuggingFace | LuaTorch

## EDUCATION

### PhD

**École Polytechnique Fédérale de Lausanne**

Mar 2021 – Ongoing

*GPA*: 5.25/6

### Integrated Masters in Information Technology | Spl. in Data Sciences

**International Institute of Information Technology, Bangalore**

Aug 2012 – Jun 2017

*Thesis Title*: Analysis of Table Tennis Strokes from RGB videos
*GPA*: 3.4/4

## PUBLICATIONS

- *Posterior-based analysis of spatio-temporal features for Sign Language Assessment*. **Tarigopula, Neha** & Tornay, Sandrine & Mercanoglu, Ozge & Bowden, Richard & Magimai-Doss, Mathew. (ICASSP/OJSP 2025)
- *Content-based objective evaluation of artificially generated sign language videos*. **Tarigopula, N.**, Garg, P., Muralidhar S., Tornay S., Jayagopi, D., & Magimai-Doss, M. (ICASSP 2024)
- *Web SMILE demo: a web application providing automated feedback on sign language vocabulary production.* Tornay, S., Nanchen, A., Battisti, A., Holzknecht, F., **Tarigopula, N.**, Mendez Maldonado, O., Camgöz, N. C., Razavi, M., Tissi, K., Sidler-Miserez, S., Boyes Bream, P., Ebling, S., Haug, T., Bowden, R., & Magimai-Doss, M. (LTRC 2023).
- *Towards accessible sign language assessment and learning*. **Tarigopula, N.**, Tornay, S., Muralidhar, S., & Magimai Doss, M. (ICMI 2022).

## ACADEMIC DUTIES

- Fall 2023: Mentored a Masters student thesis
- Spring 2022: Teaching Assistant at EPFL for Masters course in Deep Learning
- Fall 2020: Teaching Assistant at IIT Hyderabad for Bachelors course in Deep Learning for Computer Vision

## LANGUAGES

**English** Fluent
**French** Basic