THESIS

Explainable Phonology-based Approach for Sign Language Recognition and Assessment

FEBRUARY 2020

Sandrine Tornay
Idiap Research Institute
Rue Marconi 19
CH-1920 Martigny
Switzerland
sandrine.tornay@idiap.ch

Submitted to:
Doctoral school of electrical engineering (EDEE)
École polytechnique fédérale de Lausanne (EPFL)

Director : Professor **Daniel GATICA-PEREZ**
Co-director : Dr **Mathew MAGIMAI DOSS**

The more we study,
the more we discover our ignorance.
— Percy Bysshe Shelley

To my parents, my husband, my daughter, my family…

Acknowledgements

First of all, I would like to thank my main supervisor, the co-director of the thesis, Dr Mathew Magimai.-Doss. Moving from the theoretical world of mathematics to applied world was not an obvious step for me. Mathew has accompanied me perfectly well on this path. His knowledge, his clear explanations but above all his general and intuitive vision allowed me to develop myself in a broad way. Following his perfectionism, which I mainly discovered while writing articles, required me intensive work and I thank him for that because he showed me that *the more we are able to explain in a straightforward manner the more we understand the problem*. I will remember it for long. I also particularly thank him for his availability, his support and his trust. Thank you Mathew.

The first time I went to the Idiap research institute was during its 20th anniversary event. I remember thinking that working there would be a privilege. At that time, I was a student in mathematics, neurosciences and psychology, I discovered there a place which connected these fields. It was through an internship and my Master's thesis that I entered it. I would like to thank Dr. Milos Cernak who was my first supervisor at Idiap. Learning by his side brought me a lot and was the beginning of my path in the application of mathematics. It was after a detour in the world of education that I returned to Idiap to do the present thesis. I would like to thank Prof. Hervé Bourlard for providing the academic opportunity of doing a thesis jointly with EPFL. I am also very much grateful to Prof. Daniel Gatica-Perez, the director of the thesis, for his constructive feedback. I also thank the administrative team of Idiap, in particular Mrs Nadine Rousseau, Mrs Sylvie Meier and Mrs Laura Coppey for all their help, support and smiles, and the Idiap IT team for their prompt help. I also thank the speech group in which I discovered interesting topics. I would like also to extend my thanks to the committee members who evaluated this work, Dr Sarah Ebling, Dr Dinesh Jayagopi and Prof. Jean-Philippe Thiran. I am very appreciative of their valuable feedback and suggestions which helped me in improving this thesis.

This thesis was funded by the SNSF through the Sinergia project SMILE *(Scalable Multimodal Sign Language Technology for Sign Language Learning and Assessment)*, grant agreement CR-SII2_160811. Through the SMILE project, I discovered a collaborative work with people from different fields. I have learned a lot from this, both professionally and personally. Thanks to

this project, I was able to discover the world of the Deaf community which is very rich. I was fortunate to learn the first level of Swiss French Sign Language. I thank all the members of the SMILE project for the friendly meetings, the interesting discussions and pleasant stays: Prof. Dr Tobias Haug, Dr Sarah Ebling, Dr Penny Boyes Braem, Sandra Sidler-Miserez and Katja Tissi for the university of applied sciences in special needs education in Zürich, Prof Richard Bowden, Dr Simon Hadfield, Dr Oscar Mendez Maldonado, Dr Necati Cihan Camgöz and Stephanie Stoll for the university of Surrey in Great Britain, Dr Oya Aran and I particularly thank Dr Marzieh Razavi, my colleague at Idiap, for her guidance, her support and her kindness. I have been fortunate to work with her.

Doing this thesis was also a discovery of myself. Frustration, pride, demotivation, enthusiasm, stress, excitement, joy, … so many emotions to manage as well as possible. I thank the people of Idiap who, through a smile, a discussion, a break, have comforted me in my work. I would like to particularly thank Rémy for his sincere listening, our philosophical sharings and his positive energy; Noémie for her support, our valuable discussions, and her sincerity; Emmanuel for his contagious enthusiasm in learning and his optimism; Pierre-Edouard, Christian, Olivia, Hakan, David, Phil, Sylvain, Julian, Pavan, Angelos, Angel, Nicolas, Bozo, Thibault, Hannah, Alexandre for the nice discussions and coffee breaks. I have met lovely people at Idiap and I apologize for not mentioning everyone. I also take the opportunity to thank my university friends, Dan and Florian, for their friendship and our joyful moments.

From a young age, I wanted to learn mathematics. I remember the day I told my parents that I was choosing the path of study for discovering the mathematics. "Understanding those and what surrounds us" is a sentence that sums up well my motivation. I am extremely fortunate for having parents who supported me with their unconditional love and their blessings and who never doubted my abilities. My gratitude goes also to my brothers for our precious complicity, a unique support.
My heartfelt acknowledgment goes to my husband for always supporting me, for encouraging me during the tough times, and for his beautiful love. I also thank the life which allowed me to become a mother during this thesis and to give birth to our little Magalie, my everyday sunshine.

Vernayaz, February 28, 2021 Sandrine Tornay

Abstract

Sign language technology, unlike spoken language technology, is an emerging area of research. Sign language technologies can help in bridging the gap between the Deaf community and the hearing community. One such computer-aided technology is sign language learning technology. To build such a technology, there is a need for sign language technologies that can assess sign production of learners in a linguistically valid manner. Such a technology is yet to emerge. This thesis is a step towards that, where we aim to develop an "explainable" sign language assessment framework. Development of such a framework has some fundamental open research questions: (a) how to effectively model hand movement channel? (b) how to model the multiple channels inherent in sign language? and (c) how to assess sign language at different linguistic levels?

The present thesis addresses those open research questions by: (a) development of a hidden Markov model (HMM) based approach that, given only pairwise comparison between signs, derives hand movement subunits that are sharable across sign languages and domains; (b) development of phonology-based approaches, inspired from modeling of articulatory features in speech processing, to model the multichannel information inherent in sign languages in the framework of HMM, and validating it through monolingual, cross-lingual and multilingual sign language recognition studies; and (c) development of a phonology-based sign language assessment approach that can assess in an integrated manner a produced sign at two different levels, namely, lexeme level (i.e., whether the sign production is targeting the correct sign or not) and at form level (i.e. whether the handshape production and the hand movement production is correct or not), and validating it on the linguistically annotated Swiss German Sign Language database SMILE.

Keywords Sign language assessment, sign language recognition, sign language verification, lexeme-level assessment, form-level assessment, hand movement subunits, phonology-based sign language processing, hidden Markov model

Résumé

Le domaine technologique de la langue des signes, contrairement à celui de la langue parlée, est un domaine de recherche émergent. Les technologies liées à la langue des signes peuvent aider à combler l'écart entre la communauté sourde et la communauté entendante. L'une de ces technologies d'assistance est celle qui concerne l'apprentissage de la langue des signes. Pour construire une telle technologie, il est nécessaire de disposer de méthodes capables d'évaluer la production de signes d'apprenants d'une manière linguistiquement valide. Une telle technologie n'a pas encore vu le jour. Cette thèse est une étape vers cela, où nous visons à développer un cadre d'évaluation "explicable" de la langue des signes. L'élaboration d'un tel cadre comporte les questions de recherche fondamentales ouvertes suivantes : (a) comment modéliser efficacement le canal d'information du mouvement de la main ? (b) comment modéliser les différents canaux d'information inhérents à la langue des signes ? et (c) comment évaluer la langue des signes à différents niveaux linguistiques ?

La présente thèse aborde ces questions de recherche avec : (a) le développement d'une approche basée sur les modèles de Markov cachés (HMM) qui, en utilisant seulement la comparaison par paires des signes, dérive des sous-unités du mouvement de la main qui ont la propriété d'être transferable entre les différentes langues des signes ; (b) le développement d'approches basées sur la phonologie, inspirées de la modélisation des caractéristiques articulatoires du traitement de la parole, pour modéliser l'information multicanal inhérente aux langues des signes dans le cadre des HMM, et la valider par des études de reconnaissance de la langue des signes monolingues et multilingues ; et (c) le développement d'une approche d'évaluation de la langue des signes basée sur la phonologie qui peut évaluer un signe produit de manière intégrée sur deux niveaux différents, à savoir, au niveau du lexème (c.-à-d. si la production du signe vise le bon signe ou non) et au niveau de la forme (c.-à-d. si la production de la forme de la main et la production du mouvement de la main sont correctes ou non), et la valider avec la base de données SMILE de la langue des signes suisse allemande qui est annotée linguistiquement.

Mots clés Evaluation de la langue des signes, reconnaissance de la langue des signes, vérification de la langue des signes, évaluation au niveau du lexème, évaluation au niveau de la forme, sous-unités du mouvement de la main, traitement de la langue des signes basé sur la phonologie, modèle de Markov caché

Zusammenfassung

Anders als die Lautsprachtechnologie stellt die Gebärdensprachtechnologie immer noch ein junges Forschungsgebiet dar. Gebärdensprachtechnologien sind in der Lage, die Kommunikationslücke zwischen der Gehörlosen- und der hörenden Gemeinschaft überbrücken. Eine derartige assistive Technologie ist die Gebärdensprachlerntechnologie. Um eine solche Technologie zu entwickeln, werden wiederum Technologien benötigt, die Produktionen von Gebärdensprachlernerinnen in einer linguistisch validen Weise überprüfen können. Eine solche Technologie existiert noch nicht. Die vorliegende Dissertation stellt einen Schritt in diese Richtung dar, indem sie ein Framework für „erklärbare“ Gebärdensprachüberprüfung bereitstellt. Die Entwicklung eines solchen Frameworks geht mit einigen fundamentalen offenen Forschungsfragen einher: (a) Wie lässt sich Handbewegung modellieren? (b) Wie lassen sie die verschiedenen Kanäle, die der Gebärdensprache inhärent sind, modellieren? (c) Wie lässt sich Gebärdensprache auf verschiedenen linguistischen Ebenen überprüfen?

Die vorliegende Dissertation bearbeitet diese offenen Forschungsfragen, indem sie (a) einen auf Hidden-Markov-Modellen (HMM) basierten Ansatz entwickelt, der Handbewegungs-Subunits aus paarweisen Produktionen von Gebärden ableitet, die sich auf unterschiedliche Gebärdensprachen und Domänen generalisieren lassen; (b) phonologie-basierte Ansätze entwickelt, die von der Modellierung artikulatorischer Merkmale in der automatischen Verarbeitung gesprochener Sprache inspiriert sind, um die Mehrebeneninformation von Gebärdensprachen innerhalb des HMM-Frameworks zu modellieren und sie durch monolinguale, crosslinguale und multilinguale Gebärdenspracherkennungsstudien zu validieren; und (c) indem sie einen phonologie-basierten Gebärdensprachüberprüfungsansatz entwickelt, der eine produzierte Gebärde in integrativer Manier auf zwei verschiedenen Ebenen überprüfen kann, einerseits der Lexemebene (d.h. ob eine Gebärdenproduktion die richtige Gebärde beinhaltet oder nicht) und anderseits der Formebene (d.h. ob die Handformproduktion und Handbewegungsproduktion korrekt ist oder nicht), und diesen auf dem linguistisch annotierten DSGS-Datensatz SMILE validiert.

Keywords Gebärdensprachüberprüfung, Gebärdenspracherkennung, Gebärdensprachverifikation, Lexem-Überprüfung, Form-Überprüfung, Handbewegungs-Subunits, phonologie-basierte Gebärdensprachverarbeitung, Hidden-Markov-Modell

Contents

List of Figures

List of Tables

List of Acronyms

ANN	Artificial Neural Network
AF	"Articulatory" Features
ASL	American Sign Language
ASR	Automatic Speech Recognition
BSL	British Sign Language
CNN	Convolutional Neural Network
DGS	German Sign Language (Deutsch Gebärdensprache)
DSGS	Swiss German Sign Language (Deutschschweizer Gebärdensprache)
DTW	Dynamic Time Warping
GMM	Gaussian Mixture Models
HamNoSys	Hambourg Notation System
HMM	Hidden Markov Model
HMM/ANN	Hidden Markov Model / Artificial Neural Network
HMM/GMM	Hidden Markov Model / Gaussian Mixture Models
KL	Kullback-Leibler
KL-HMM	Kullback-Leibler divergence-based Hidden Markov Model
KLT	Kahunen Loeve Transform
LQR	Linear-Quadratic Regulator
LSF	French Sign Language (Langue des Signes Française)

LSTM	Long Short Term Memory
MLP	Multilayer Perceptron
PLP	Perceptual Linear Prediction
PRR	Phone Recognition Rate
RA	Recognition Accuracy
RNN	Recurrent Neural Network
SiGML	Signing Gesture Markup Language
SKL	Symmetric Kullback-Leibler
SLR	Sign Language Recognition
TSL	Turkish Sign Language
TTS	Text-to-speech Synthesis

1 Introduction

Humans are social beings. They like to be surrounded by friends and share their personal experiences with others. Most of us at some point would have had the frustrating experience of not being able to communicate with a person who does not understand our language. In such circumstances, in the hearing community when we cannot use the spoken language to share information, gestures come into play automatically, as if it is a universal understanding. In comparison, in the Deaf [1] community when the sign language is not known, they would use naturally mimes to communicate. Sign languages are not just pantomime but they have evolved into a language with its own vocabulary and grammar. This evolution is culturally dependent on the place where it belongs to; sign language is not universal as we may think. Various tools are available to acquire new spoken language such as courses, books, learning platforms, most of the methods allow to learn it in a self-taught manner. Moreover, the correspondence between written and spoken language allows a "dual learning". Learning as a whole is more accessible in the hearing community from the fact that the written form of the known language facilitates the self-taught learning. In the Deaf community, the accessibility of learning is another story since the written language linked to the sign language is similar to a foreign language. Thus, computer-aided tools are needed as support. In other words, automatic sign language processing can help in bridging the gap between the hearing and the Deaf community by developing recognition systems, machines translation, or learning platforms.

1.1 Goal of the Thesis

The goal of the thesis is to develop a framework which allows to assess isolated sign production in sign language. As depicted in Figure 1.1, a framework that is not only able to verify the

[1] The upper-cased word Deaf is conventionally used to describe the members of the linguistic community of sign language users and, in contrast, the lower-cased deaf to describe the audiological state of a hearing loss [75]

produced sign in a linguistically valid manner, i.e whether the produced sign is acceptable or not (referred to as sign verification), but also provides linguistically guided feedback on the production of the sign, i.e. detailed analysis based on linguistic annotations. In this thesis, we refer to sign verification together with detailed analysis based on linguistic annotation as sign language assessment.

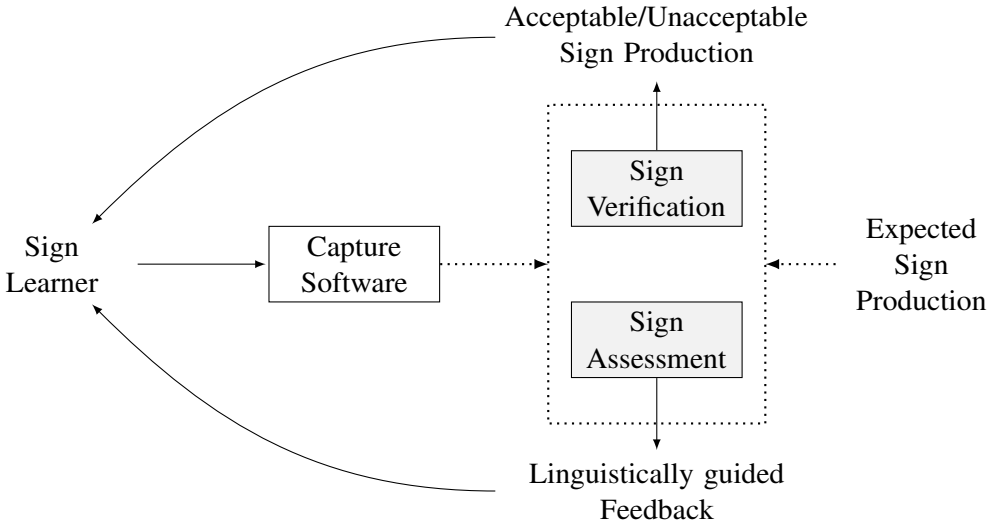

Figure 1.1 – Illustration of the sign language assessment framework which allows to verify a produced sign in a linguistically valid manner and provides linguistically guided feedback on the production of the sign.

Three main challenges are encountered in that direction:

1. Resource scarcity: Only few sign languages have a proper database. Also, unlike spoken languages, sign languages users are limited. Furthermore, sign languages have their own vocabulary and grammar, different than the corresponding spoken language [109]. For instance, British Sign Language is not a signed form of British English. Furthermore, even though the spoken language can be the same, the sign languages can be different. For example, American Sign Language and British Sign Language are different sign languages. Similarly, Swiss German Sign Language and German Sign Language are different sign languages. As a consequence, it is not trivial to share resources from different sign languages.

2. Sign language linguistics: Even if there are enough resources, advances are still needed in sign language linguistics. Few formal sign language reference grammars exist. Sign assessment is not only about gesture verification: linguistic constraints given by the syntax rules restrict the combination of movements, so they have to be considered. Towards

that, the definition of the correctness, i.e. which deviations are acceptable or not, has to be defined. This information is not readily available. Indeed, this aspect is still a point of research in the sign linguistics community [34]. Another relevant linguistic field is the phonology. As stated in [16], "Sign language phonology is the abstract grammatical component where primitive structural units are combined to create an infinite number of meaningful utterances." Such knowledge about minimal units, such as phonemes / graphemes in spoken/written language, is valuable information in sign language processing such as to derive the structure of the sign's model.

3. Sign language processing technologies: To convey information, sign languages use simultaneously manual and non-manual channels of information such as the handshape, hand movement/position/location/orientation, the facial expression, mouthing, the movement of the torso. First, the relevant information of these channels has to be properly extracted, which is itself a challenge. Then, the sign language recognition model not only has to integrate the multichannel aspect of the sign, but it has to be explainable. Indeed, in the assessment task, we want it to be sensitive to the production variation of each channel and transparent, i.e. the production information can be fed back spatially (which channel) and temporally (which time frame), for providing detailed feedback. Such an explainable framework is yet to emerge in the sign language technology community.

To address these challenges, a collaborative work is needed. The development of the assessment framework of this thesis took place in the context of the SNSF Sinergia project SMILE[2]. Broadly, the goal of the SMILE project was to develop an advanced platform which allows to assess Swiss German Sign Language (Deutschschweizer Gebärdensprache) (DSGS). This project used a multidisciplinary framework which involved three complementary partners:

- The **Hochschule für Heilpädagogik** (HfH) in Zürich brought its expertise in sign language linguistics and assessment disciplines. Specifically, they developed the DSGS resources by collecting the data, providing data transcription, assessing the acceptability of the sign, annotating the error on the production, as well as the overall structure of the assessment framework.

- The **University of Surrey** (USurrey) in UK brought its expertise in sign language technology and computer vision. Specifically, they participated in the DSGS data collection. They developed the recording software, the capture tools, the data acquisition methods. They also developed the handshape estimator and worked on the automatic production/synthesis of sign language.

[2]SMILE stands for Scalable Multimodal sign language technology for sIgn language Learning and assessmEnt; http://www.idiap.ch/en/scientific-research/projects/SMILE

- The **Idiap Research Institute** (Idiap) in Martigny, Switzerland, brought its expertise in Hidden Markov Model (HMM) applied to pronunciation generation, pronunciation modeling and speech recognition. Specifically, we developed approaches to extract and model hand movement subunits and the framework to integrate the handshape and the hand movement channels for sign language recognition and assessment.

1.2 Contributions of the Thesis

The main contributions of the thesis are:

1. The development of approaches to model hand movement as discrete units. State-of-the-art neural networks-based sign language recognition methods focused on the handshape channel mainly thanks to the discrete aspect of the handshapes which makes available the transcription of the sign into them. But it is no sufficient, the handshape and the hand movement are used jointly to convey sign's meaning. Hand movement information is continuous in nature. As a consequence, representing and modeling the hand movement information in sign language production as a sequence of discrete units is not a trivial task. In that direction, we develop methods based on HMM that, using position and velocity features extracted from the visual signal, gives discrete symbolic representation of hand movement. This representation gives an alternative to handle the challenges related to: (a) effective modeling of hand movement information along with handshape information for sign language recognition and sign language assessment and (b) addressing of data scarcity issues.

   The following publications are part of the first contribution:

> *An HMM approach with inherent model selection for sign language and gesture recognition*, Sandrine Tornay, Oya Aran and Mathew Magimai.-Doss, in: Proceedings of the International Conference on Language Resources and Evaluation LREC 2020, 2020

> *Subunits inference and lexicon development based on pairwise comparison of utterances and signs*, Sandrine Tornay and Mathew Magimai.-Doss, in: Information, Special Issue: Computational Linguistics for Low-Resource Languages, 10:298, 2019

Data-driven movement subunit extraction from skeleton information for modeling signs and gestures, Sandrine Tornay, Marzieh Razavi and Mathew Magimai.-Doss, Research Report, Idiap-RR-02-2019

2. The development of a phonology-based sign language recognition framework that integrates the multichannel aspect of a sign. In the past, this problem has been attempted through use of sensors such as gloves or accelerometer for modeling hand gestures. As mentioned earlier, sign language is more than hand gestures. Extracting the multichannel information from the visual signal and modeling them jointly is still an open research problem. In that direction, inspired from the works on modeling speech production information in speech processing, this thesis proposes novel HMM-based frameworks that allow extraction of each channel from the visual signal in a separate manner and jointly model them for sign language processing.

   The following publications are part of the second contribution:

HMM-based approaches to model multichannel information in sign language inspired from articulatory features-based speech processing, Sandrine Tornay, Marzieh Razavi, Necati Cihan Camgoz, Richard Bowden and Mathew Magimai.-Doss, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019

Towards multilingual sign language recognition, Sandrine Tornay, Marzieh Razavi and Mathew Magimai.-Doss, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020

3. The development of a linguistically valid sign language assessment framework. Emerging sign language assessment studies are primarily based on sign language verification [112, 49]. It is not clear whether it was done in a linguistically valid manner or not. As discussed earlier, development of an approach that is linguistically valid, first requires linguistic knowledge about which variations in sign language production are acceptable or not. Next, development of such an approach also needs data with such linguistic annotations. Third, it requires a methodology that can tease out the variations in the different channels w.r.t "expected" sign production in an explainable manner. The third main contribution builds upon the developments from the second contribution to propose a sign language assessment framework that can effectively assess sign production in a linguistically valid manner and provide linguistically guided feedback.

The following publication is part of the third contribution:

> *A phonology-based approach for isolated sign production assessment in sign language*, Sandrine Tornay, Necati Cihan Camgoz, Richard Bowden and Mathew Magimai.-Doss, in: Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion), 2020

1.3 Organization of the Thesis

Below is the organisation of the remainder of the thesis:

Chapter 2 provides an introduction to sign languages followed by a state-of-the-art overview of sign language processing. The databases and the evaluation measures used in this thesis are defined in this chapter.

Chapter 3 presents the manual feature estimators and the statistical sequence modeling methods used in this thesis.

Chapter 4 focuses on the temporal structure of the sign: *how to determine the appropriate HMM topology for modeling the hand movement information?*. It is a part of the first contribution (see Section 1.2) where we present a data-driven model selection approach which determines, for each sign, the appropriate HMM topology at the time of recognition, as opposed to pre-setting it. Sign language recognition and gesture recognition studies are presented to validate the proposed data-driven model selection approach.

Chapter 5 focuses on *how to discretize and model hand movement information as subunits?*. This chapter presents the core part of the first contribution (see Section 1.2) where the development of the skeleton information based hand movement subunits is presented. The proposed approach is validated through monolingual and cross-lingual sign language recognition studies and through analysis of subunits through synthesis of hand movement information.

Chapter 6 focuses on integration of the multichannel aspect of the sign production, i.e. *how to model the multichannel information inherent in sign languages?*. It presents the works related to the second contribution (see Section 1.2). We propose two HMM-based approaches to model the multichannel information. We validate those approaches through joint modeling of hand movement information and handshape information, and conducting monolingual, cross-lingual and multilingual sign language recognition studies.

Chapter 7 focuses on development of sign language assessment framework, i.e. *how to assess*

isolated sign productions at the lexeme-level (whether the produced sign is targeting the correct sign or not) and the form-level (whether the produced hand movement and handshape are correct or not)?. It presents the work of the third contribution (see Section 1.2) where we propose an explainable phonology-based sign language assessment system that builds on one of the HMM-based frameworks developed in Chapter 6. We validate the proposed approach on isolated signs using a linguistically annotated DSGS corpora developed as part of the SMILE project.

Chapter 8 finally concludes the thesis with suggestions for possible directions for future research.

2 Background

This chapter is organized as follows. Section 2.1 first introduces briefly sign languages. Section 2.2 then presents a concise overview on sign language data acquisitions and sign language recognition. Section 2.3 and Section 2.4 present the sign language databases used in the present thesis and the evaluation metrics used to evaluate different sign language processing systems, respectively.

2.1 Sign Languages

As mentioned in the introduction (see Chapter 1), sign languages are languages as complex as any spoken languages despite the common misconception that it is a universal communication composed of mimes. It is related to the spoken language of the place, but it is not a word-by-word conversion of it. Sign language is an independent language whose grammar is different from spoken/written grammar. The 2020 edition of the *Ethnologue - Languages of the World* lists 144 sign languages: 125 Deaf community sign languages and 18 shared-signing languages; the second being the sign languages developed in a shared community with hearing and Deaf members also called "village sign languages" [77].

Sign language is not solely about hand gestures. In addition to manual activity, sign language also uses shoulders/torso, head, facial expression and mouthing to convey meaning. For example, in Swiss LSF, the movement of the torso point out the tense of the sentence: forward for the future, straight for the present and backward for the past [2]. Thus, sign language interpretation requires integration of parallel multichannel information in comparison to spoken languages which typically deals with modeling of sequence of acoustic feature vectors (i.e., single stream of information). The multichannel framework can be broken down into visual subunits, sometimes called visemes, in much the same way as words can be broken into phonemes and articulatory features in spoken language. Sign language phonology deals with the study of the basic articulator

units used to produce the lexical entities such as, words. The visual subunits can be grouped into two categories: manuals, which deal with handshape, hand placement, movement, orientation and arrangement [105]; and, non-manuals, which deal with body posture, movement of the head and shoulder, facial expression with eyebrow, mouth, eyes, and cheeks. Commonly, interpretation deduces that the upper part of the face is used to mark sentence type while the lower part of the face serves adjectival or adverbial function [24]. As we all have a favourite hand when producing hand gestures, there is a distinction between the right-handed and the left-handed people. The left-handed signers mirror the right-handed signs and vice versa. Thus, the main hand used to sign is called the dominant hand and the other the non-dominant hand. Sign language is produced in a limited space, called signing space [2], which is a rectangular region that includes signer hips up to the head, as illustrated in Figure 2.1.

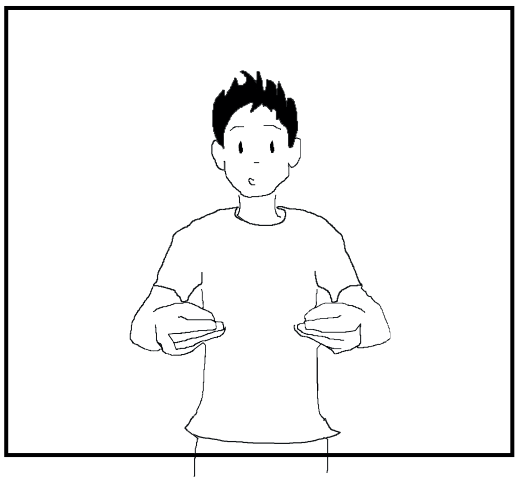

Figure 2.1 – Illustration of the signing space of a signer.

There are two main written forms of sign languages: glosses and HamNoSys.

- Glosses provide semantic labels of signs. A gloss of a sign is the written form of the most closely corresponding spoken language. It is written in upper case letters to differentiate it from the written form of the spoken word. For example, the gloss VOLK is used to represent the Swiss German Sign Language sign for 'folk', see caption of Figure 2.2. Sometimes a unique word is not sufficient because there exists no exact translation; in that case several words are joined by hyphen. Additional symbols are used to complement, for example '+++' to express repetition. To represent a finger spelled word, the gloss is written with small letters with a hyphen in between each letter.
- HamNoSys[1] is a notation system developed over 25 years ago. HamNoSys has approxi-

[1] https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/hamnosys-97.html (visited on 02/19/2021)

mately 200 symbols. It has production symbols to define the handshape, the orientation (finger direction, palm orientation), the hand location and the hand movement. For example, Figure 2.2 shows the HamNoSys annotation of the gloss VOLK and its corresponding sign production. It is focused on the description of manual activities. It is one of the most developed annotation for sign language. Compared to phonetic notation of spoken language which is mainly perceptive notation, HamNoSys is both perceptive and production based but it does not tell about how human perceives the multiple channels of information (manuals and non-manuals) jointly. From HamNoSys, what we can deduce is that the information is transmitted by spatial and multichannel parallel linear arrangement in a temporal dimension. To overcome the temporal deficiency of HamNoSys annotation, there exists the SiGML [37] annotation which is an XML representation of HamNoSys.

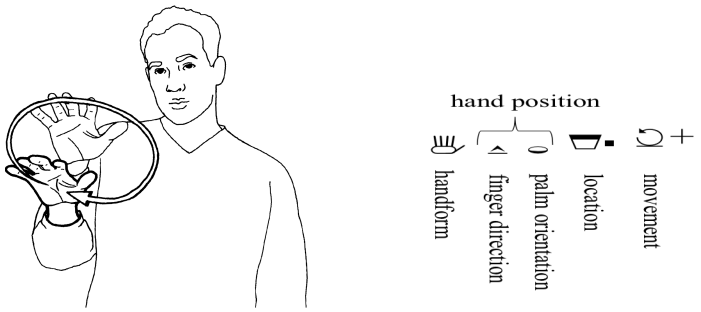

Figure 2.2 – HamNoSys annotation for VOLK, the Swiss German sign for 'folk'.

2.2 Sign Language Processing

In this section, we provide a concise survey on the data acquisition techniques and the sign language processing techniques. Sign language processing research can be split into three major directions: Sign Language Recognition (SLR), sign language production/synthesis and sign language translation. This thesis mainly concerns with SLR this is why the sign language processing presentation limits to SLR. For an extensive overview on sign language processing, the reader is referred to [28].

2.2.1 Data acquisition

Sign language data acquisition methods have evolved over the years. The acquisition methods can be split in two categories: sensor-based and camera-based. To capture users' hands and body movements, early SLR studies have used sensor-based approaches such as data gloves [118, 96] or colour gloves [103, 102, 104, 6, 48]. In recent years, gloves have been designed to translate sign language into spoken language [26, 97]. However, as discussed earlier, even if the

manual components are the major part of sign language, the non-manual aspect is key parts of the grammar especially for continuous sign language.

In the camera-based case, there are a few studies using 2D features with tracking [129, 10]. However, 2D visual information modeling lacks the depth component, which is an important component of sign language. Furthermore, occlusions (hand-to-face or hand-to-hand) cannot be handled. To overcome these limitations, acquisition of 3D information using multiple cameras [115] or with calibrated light source [99] have been proposed. The release of depth cameras, such as the Microsoft Kinect sensor [132], allows to decrease the dependency on multiple sensors, and facilitates the acquisition of 3D components by providing depth map and real-time human pose information [100]. For further readings on data acquisition and visual feature extraction, the reader is referred to [35, 28].

2.2.2 Sign language recognition techniques

Sign language recognition techniques in the literature can be grouped into three main categories: Artificial Neural Network (ANN)-based, HMM-based and other pattern recognition techniques based.

Artificial neural networks and variants

The earliest works on SLR applied ANN. ANN were often used to classify particular visual subunit of the sign separately as handshape, hand location, orientation or to track hand movement. Most of these methods focused on isolated sign classification. One of the first SLR work [76] modelled data acquired through gloves (finger and hand angles, hand position) with a Recurrent Neural Network (RNN). The RNN could take into account temporal aspect, but it failed to address segmentation of the signs across time. Waldron and Kim [118] combined gloves information with a tracking sensor. They first trained an ANN for each of the four subunit types (handshape, hand location, orientation and movements), and then combined the outputs of the different subunits ANN with a second ANN that classifies isolated signs. This method allows to separately model multichannel information with the ANN in the first stage but the second ANN does not yet allow to segment the produced sign in time. Yang et al. [127] proposed a temporal processing method that models 2D hand pixel motion trajectories using time-delay neural network to classify signs.

More recent works have focused on using Convolutional Neural Network (CNN) to extract handshape features by directly feeding with cropped hand images. Pigou et al. [88] have used such an approach to feed an ANN classify isolated signs. Oliveira et al. [78] have compared the CNN-based approach with principal component analysis combined with the k-nearest neighbour algorithm for handshape classification. Xie et al. [125] have investigated use of CNNs and

support vector machine or SoftMax classifier for recognition of hand gestures. For continuous sign language recognition, Koller et al. [62] used a CNN to classify the handshapes based on cropped images and a HMM to model the extracted handshape information. Camgoz et al. [20] proposed to combine spatial modeling with CNN using handshape cropped images, temporal modeling with Long Short Term Memory (LSTM) and sequence-to-sequence learning with connectionist temporal classification.

Hidden Markov models and variants

Sign language recognition can be considered as a sequence recognition problem, similar to speech recognition. This similarity has led to transfer of techniques from speech processing to sign language processing. In [103], based on the use of hand gloves, the authors proposed a four-state left-to-right HMM with one skip transition for isolated sign language recognition. In [115], the authors used computer vision-based methods as well as a magnetic sensor system to extract handshape and hand movement. They addressed the co-articulation effects in continuous signing and model the transition movements between signs and within the signs themselves with different HMM topologies. One of the main limitations of these works is that the experimental studies were carried out on signer-dependent setup, i.e. the models were trained and tested on the same signer.

Different HMM variants have been developed in the literature for SLR. In [57], the authors proposed a method to model handshape features using input-output HMM. One of the problems in HMM is to model the parallel multichannel information inherent in signs. In that context, parallel HMM have been used where a separate HMM is used for modeling right and left hands or different feature sets such as handshape, hand configuration and motion [117, 131]. Vogler and Metaxas demonstrated that parallel HMM tend to yield better systems when compared to standard HMM, factorial HMM and coupled HMM [117]. With the development of Kinect sensor, Kumar et al. [66] developed a multi-sensor fusion framework for isolated sign recognition by modeling state-space dependency using coupled HMM. Liu et al. [72] demonstrated that left-to-right HMM topology to model signs provides the best performance compared to fully connected HMM topology. Another issue when using left-to-right HMM is to fix the model size. All the above discussed systems used a fixed number of states for all signs. In the literature, there have been works to set the number of states based on model selection [101, 120, 68, 59].

In addition to the above discussed approaches, there are methods that combine visual subunit classification and sequence modeling. One such approach is the two stages approach. In this approach, the first stage consists of visual subunits classifiers (handshape, placement, motion). The second stage consists of a sequential model for each sign constructed based on first order Markov assumption using the outputs of the first stage classifiers [27, 29].

The above discussed works have mainly focused on isolated sign language recognition. HMM have also been used for continuous SLR [128, 116].

Other approaches

Sign language recognition has been also approached using other pattern recognition techniques such as, Support Vector Machine [91], Dynamic Time Warping (DTW) [70], Conditional Random Fields [63], Gaussian process dynamical models [41], to mention a prominent few.

2.3 Databases

In this section, we present the databases that have been used in the thesis.

2.3.1 Chalearn14 gesture database

The Chalearn14 database was collected in the context of the ChaLearn Looking at People 2014 challenge [38] which contains 3 tracks: (1) human pose recovery, (2) action/interaction recognition, (3) multi-modal gesture recognition. In this thesis, we used the data of the third track which is based on the Italian gesture database, called Montalbano gesture database (from ChaLearn 2013 [39]). It includes a vocabulary of 20 Italian cultural/anthropological gestures. The 27 users are recorded in a wild environment in front of a Kinect, performing natural communicative gestures and speaking in fluent Italian. 81% of the participants were native Italian speakers. Each sign has been repeated several times by each user. The database is publicly available[2].

2.3.2 DGS database

The DGS database was collected to study wide variety of signing styles [79]. It contains 40 signs from German Sign Language (DGS). The database includes data from 14 non-native right-handed signers, where each sign is repeated approximately 5 times by each person. There are a total of 3186 signs in the database. The DGS is a challenging database as the signs performed by the non-native signers contain large variation. The database has been recorded in an uncontrolled environment with a Kinect camera. The 3D coordinates of a human skeleton has been tracked using the OpenNI framework [81].

[2] http://gesture.chalearn.org/mmdata#Track3 (visited on 02/19/2021)

2.3.3 HospiSign database

The HospiSign database is the subset of the BosphorusSign database [19] which contains the signs related to the health domain. These data were collected in the context of developing the HospiSign [110] interactive sign language interface for hospitals. The subset contains 33 phrase classes from Turkish Sign Language (TSL). The HospiSign subset includes six signers signing each phrase approximately 6 times [18]. The database has been recorded with a Kinect camera. The database is publicly available by request from the authors[3].

2.3.4 SMILE DSGS database

The large-scale SMILE DSGS database [34] was created in the context of developing an assessment system for lexical signs of Swiss German Sign Language (DSGS) in the SMILE project. It contains 11 adult L1 signers and 19 adult L2 learners who produced 100 isolated signs of a DSGS vocabulary production test. The 100 lexical items were chosen based on learning material of the A1 DSGS level (see [34] for more details). Each sign was performed three times and only the second pass was manually annotated. The SMILE DSGS database was collected with the Microsoft Kinect v2 sensor and the high speed and high resolution GoPro video cameras. The SMILE DSGS database provides the colour videos, depth maps, user masks and 3D pose information obtained from the Kinect, the body pose, facial landmarks, and hand pose information extracted using the deep-learning-based key point detection library OpenPose [21].

In this thesis, we only used the second pass annotated data with the 'Category of sign produced' annotation of the SMILE transcription/annotation scheme (presented in detail in [34]). Briefly, this linguistic annotation evaluates, through six categories, the acceptability of a sign according to three linguistic criteria: lexeme, meaning and form. Table 2.1 presents the different categories and the distinction between them.

Table 2.1 – SMILE annotation scheme of the 'Category of sign produced' annotation presented in [34]

Category	Same lexeme as target sign?	Same meaning as target sign?	Same form as target sign?
cat.1	yes	yes	yes
cat.2	yes	yes	slightly different
cat.3	yes	yes	no
cat.4	yes	slightly different	slightly different
cat.5	no	yes	no
cat.6	no	no	no

[3] https://www.cmpe.boun.edu.tr/pilab/BosphorusSign/home_en.html (visited on 02/19/2021)

We did not make any difference between the L1 and L2 signers in our experiments. To ensure that enough samples are available for each sign (minimum 5 samples/sign), 94 signs were selected out of the 100.

2.4 Evaluation Measures

Two tasks are addressed in this thesis: (i) isolate SLR and (ii) isolated sign language assessment. In the SLR task, the recognition accuracy (RA) is defined as the number of correct predictions divided by the total number of samples, i.e.

$$\text{RA} = \frac{\#\ of\ correctly\ predicted\ signs/gestures}{total\ \#\ of\ signs/gestures\ in\ the\ reference}\ . \tag{2.1}$$

The sign language assessment task is a detection task. So, we used the F_1 score as the evaluation measure. The F_1 score is the harmonic mean of the precision (denoted as p) and the recall (r), i.e.

$$F_1 = 2 \cdot \frac{p \cdot r}{p + r}\ , \tag{2.2}$$

where $p = \frac{TP}{TP+FP}$ and $r = \frac{TP}{TP+FN}$; TP stands for true positives, FP for false positives and FN for false negatives.

2.5 Summary

In this chapter, we first provided a brief overview on sign languages. We then presented different sign language data acquisition methods and different approaches for sign language recognition based on ANN and HMM. Finally, we described the four different databases used in this thesis and the evaluation measures used for SLR task and sign language assessment task.

3 Features and Statistical Modeling Methods

This chapter presents in Section 3.1 the extraction of the manual features used in this thesis, namely the hand movement and the handshape feature extractors. Then, in Section 3.2 and Section 3.3, a background on the statistical sequence modeling techniques used in this thesis is provided. Finally, Section 3.4 concludes with a summary of this chapter.

3.1 Manual Features

As described earlier in Section 1.1, to convey meaning, sign language uses hand gestures but also non-manual components such as facial expression, body posture or lip movement. While non-manual components are more used for complementary information and grammar (in continuous sign language) such as the interrogative form, hands focus on the semantic meaning. This explains why SLR models focus mainly on manual features. There exist two linguistic-oriented approaches which explain these syntax rules: the Stokoe system [105] and the Movement-Hold model [71]. In the Stokoe system, a sign is described as a simultaneous series of three major formational units: handshape, hand locations and hand movements, while the Movement-Hold model fragments the signs in two types of sequentially ordered segments: movement and hold (location) segments.

In this thesis, we focus only on manual features to model isolated signs, phrases and gestures. One of the main reasons being extraction of non-manual features from the visual signal in a systematic manner is still an open topic for research [28]. We separate the manual features into two main channels: the hand movement features and the handshape features. These features are extracted from the visual signal captured using Microsoft Kinect sensor (see Section 2.3). Throughout the thesis, for consistent modeling reasons, the dominant hand is treated as the right hand and mirror operation is applied to left-handed signers.

3.1.1 Hand movement features

To model the hand movement information of a sign, we focused on two components: the hand locations and the hand trajectories. For sake of simplicity, in the remainder of the thesis, we use the term *hand movement* to refer to them. To model it, we used 3D skeleton pose data. While the manual signals are the basic components that form the signs, several other key body parts such as the face, shoulders and arms are also important in the analysis of manual signs in order to understand the relative position of the hands with respect to the body [5]. Given the x, y, z coordinate of the hand, we used two parameters: the position linked to the hand location and the velocity for the hand trajectories. To represent the signer's body and to handle the variation in-between the signers, we decided to use three coordinate centers: the head, the shoulders and the hips (see Figure 3.1). According to that, position features are given by the 3D coordinate of both

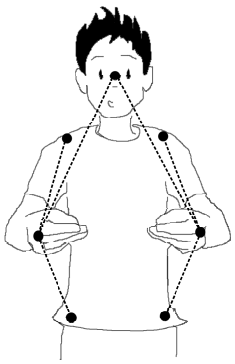

Figure 3.1 – Both hands movement features are expressed in three coordinate centers (head, shoulder and hip joint) based on x, y, z coordinates of the skeleton joints.

hands according to the head, shoulder, hip coordinate and velocity features are the delta features. More precisely, for each time frame t, we normalized position features of the non-dominant, $\mathbf{p}_t^{\text{lhnd}}$, and the dominant hand, $\mathbf{p}_t^{\text{rhnd}}$, by the width of the head, i.e.

$$\mathbf{p}_t^{\text{lhnd}} = \frac{\mathbf{lhnd}_t - \mathbf{center}_t}{|\text{neck}_{y,t} - \text{head}_{y,t}| \ /4}\ , \tag{3.1}$$

$$\mathbf{p}_t^{\text{rhnd}} = \frac{\mathbf{rhnd}_t - \mathbf{center}_t}{|\text{neck}_{y,t} - \text{head}_{y,t}| \ /4}\ , \tag{3.2}$$

where $\mathbf{lhnd}_t$, $\mathbf{rhnd}_t$ are vectors containing the x, y, z coordinates at time frame t of related joints: non-dominant hand and dominant hand; $\mathbf{center}_t$ are vectors containing the x, y, z coordinates at time frame t of related joints: head, right/left shoulder, right/left hip where the right shoulder/hip center is used to compute the dominant hand position vector and the left shoulder/hip for the non-dominant one; $\text{head}_{y,t}$ and $\text{neck}_{y,t}$ are the y coordinate of the head and the neck joint at time frame t. The related velocity features, $\mathbf{v}_t^{\text{lhnd}}$ and $\mathbf{v}_t^{\text{rhnd}}$, are estimated by computing the difference

between the position features at time t and time $t-2$, i.e.

$$\mathbf{v}_t^{\text{lhnd}} = \mathbf{p}_t^{\text{lhnd}} - \mathbf{p}_{t-2}^{\text{lhnd}} , \tag{3.3}$$

$$\mathbf{v}_t^{\text{rhnd}} = \mathbf{p}_t^{\text{rhnd}} - \mathbf{p}_{t-2}^{\text{rhnd}} . \tag{3.4}$$

Thus, after normalization, the stack of the continuous hand motion and position values related to the three coordinate centers give us the necessary information on the hand trajectory and position with respect to the signer's body.

The resulting feature vector, $\mathbf{x}^{\text{hmvt}}$, is the stack of $\mathbf{p}^{\text{lhnd}}$, $\mathbf{p}^{\text{rhnd}}$, $\mathbf{v}^{\text{lhnd}}$ and $\mathbf{v}^{\text{rhnd}}$ according to the three coordinate centers, leading to a sequence of $F \times 36$ dimensional feature vectors, where F is the total number of frames and the 36-dimensional feature vector consists of 18 position features (= 3 coordinates · 3 coordinate centers · 2 hands) and 18 velocity features.

Shoulder normalization-based features: In this thesis, we carry out cross-lingual and multilingual studies involving multiple sign language databases (see Chapter 2.3) which have different recording settings such as standing or sitting position. In order to compensate the differences in the coordinate system in-between the three databases, before feature extraction, we aligned the skeletons of signers, **skel**, irrespective of the datasets, w.r.t a chosen reference signer at the neck joint and scaled by the shoulder width. More precisely, we first expressed the skeleton joints in the neck coordinate center system, $\mathbf{skel}^{\text{neck}}$, i.e.

$$\mathbf{skel}^{\text{neck}} = \mathbf{skel} - \mathbf{neck} , \tag{3.5}$$

where **neck** are the neck joints of the input skeleton. We then computed the shoulder-to-shoulder distances, f, in-between the reference and the input skeleton used to normalize the skeleton joints, i.e.

$$f = \frac{|\, \mathbf{lshd}_{\text{ref}} - \mathbf{rshd}_{\text{ref}} \,|}{|\, \mathbf{lshd} - \mathbf{rshd} \,|} , \tag{3.6}$$

where $\mathbf{lshd}_{\text{ref}}$, $\mathbf{rshd}_{\text{ref}}$ and **lshd**, **rshd** are respectively the left and the right shoulder joints of the reference and the input skeletons. We finally normalized the input skeleton aligned to the neck and translated it to the new reference neck coordinate center system, resulting to

$$\mathbf{skel}_{\text{norm}} = f \cdot \mathbf{skel}^{\text{neck}} + \mathbf{neck}_{\text{ref}} , \tag{3.7}$$

where $\mathbf{neck}_{\text{ref}}$ is the neck joint of the reference skeleton. The resulting feature vector is denoted as $\widehat{\mathbf{x}}^{\text{hmvt}}$ in the remainder of the thesis.

3.1.2 Handshape features

Figure 3.2 illustrates the process of handshape feature extraction. More precisely, cropped hand sequences of the sign is obtained from the visual signal and fed as input to an ANN to estimate handshape "subunits" posterior probabilities.

Figure 3.2 – Illustration of the handshape subunits posterior probability features estimator.

In this thesis, two different handshape estimators were used: one in the context of sign language recognition and one in the context of sign language assessment. The development of these estimators were done by USurrey partners[1] of the SMILE project (see Chapter 1.1). The handshape subunits classifiers were built using the One-Million-Hands[2] dataset, which contains cropped hand patches and the aligned handshape labels from three different sign languages, namely Danish Sign Language, New Zealand Sign Language and German Sign Language. This dataset was originally created in the context of developing a frame-based handshape classifier on weakly annotated data [60]. The handshape labels consist of a transition shape and 60 linguistically inspired handshape classes or subunits presented in detail on the following website: https://www-i6.informatik.rwth-aachen.de/~koller/1miohands-data/ (visited on 02/19/2021).

Handshape feature extraction for sign language recognition

For the sign language recognition task, we used the off-the-shelf CNN-based DeepHand approach developed by Koller et al. on the One-Million-Hands dataset [60] and available on the website mentioned earlier. For each cropped hand patch input at time frame t, the CNN estimates a 61 dimensional handshape subunits posterior probability vector $\mathbf{z}_t^{\text{hshp}}$. For the two hands, this yields two 61-dimensional vectors.

Handshape feature extraction for sign language assessment

In the sign language assessment task, the ANN classifier is based on the CNN-LSTM hybrid proposed by Camgoz et al. in [20], called SubUNets, which uses Connectionist Temporal Classification loss layer [43] to avoid the costly iterative realignment process of the DeepHand [60] approach (used in the sign language recognition task). More recent residual network based

[1] more precisely by Cihan Necati Camgoz that I particularly thank for his availability and his support.
[2] https://www-i6.informatik.rwth-aachen.de/~koller/1miohands-data/ (visited on 02/19/2021)

CNN architecture, namely ResNeXt-101 [126], was used instead of the AlexNet [64] originally presented in [20]. Moreover, the One-Million-Hands dataset is unbalanced causing the networks to learn prior over the present 60 handshape classes, which does not necessarily extend to other domains or other sign languages. To improve the quality and the generalization of our handshape subunit representations, we first reduced the number of classes by choosing the most common handshape classes present in the One-Million-Hands dataset by keeping the handshape classes which have at least 1000 samples in the training set. This reduced the number of classes to 30. We then collected new samples from four participants, two L2 signers and two non-signers, to handle the class imbalance problem. We applied random rotation, zoom and colour jitter to help our networks generalize better. To further address the class imbalance issue, we re-sampled the training images with respect to their corresponding classes and simulate a uniform distribution over all classes. Finally, the 30-dimensional handshape classifier was trained and in addition to that a second 31-dimensional handshape classifier which includes a transition shape.

The inference process starts by cropping hand patches. The SMILE DSGS dataset [34] was recorded using the Microsoft Kinect v2 depth sensor. Although Kinect SDK provides 2D joint locations for colour images, its wrist localization is too jittery. To overcome this problem, we utilized a state-of-the-art 2D pose estimation method, namely OpenPose [21], to localize wrist locations. Using the wrist pixel coordinates, the non-dominant and dominant hands patches are cropped to extract posterior probabilities of handshape subunits on both classifiers resulting in a 122-dimensional handshape posterior probability vector (= 2 hands $\cdot$ (30 + 31)). The models were implemented using the PyTorch deep learning framework [86].

3.2 Statistical Sign Language Recognition Framework

In the statistical SLR approach, given an input video as a sequence of images/features $X = (\mathbf{x}_1, \cdots, \mathbf{x}_T)$, the goal is to obtain the most likely sign (in the case of isolated SLR) or sign sequence S^* (in the case of continuous SLR), i.e.,

$$S^* = \arg\max_{S \in \mathscr{S}} P(S|X, \Theta) \ , \tag{3.8}$$

where $\mathscr{S}$ denotes the set of all possible signs or sign sequences, S represents a sign or sign sequence and Θ denotes the set of parameters of the system. For simplicity, in the remainder of this section Θ is dropped. As direct estimation of $P(S|X)$ is a non-trivial task[3], typically Bayes'

[3]It is worth mentioning that recently there are approaches emerging which directly model $P(S|X)$ [43, 20].

rule is applied, leading to,

$$S^* = \arg\max_{S \in \mathscr{S}} \frac{p(X|S)P(S)}{p(X)} , \tag{3.9}$$

$$= \arg\max_{S \in \mathscr{S}} p(X|S)P(S) . \tag{3.10}$$

Eqn. (3.10) is obtained as a result of the assumption that $p(X)$ does not affect the optimization. $P(S)$ is referred to as the language model, and can be estimated based on the relative frequency of the signs on the training data. A common way to model $p(X|S)$ in the literature is to use conventional HMM-based approaches [85], where HMM is a Markov process with hidden states/variables $q_1, \ldots, q_t, \ldots, q_T$. Thus, $p(X|S)$ in an HMM-based framework can be estimated by summing over all possible hidden state sequences $\mathscr{Q}$, i.e.,

$$p(X|S) = \sum_{Q \in \mathscr{Q}} p(X, Q|S) , \tag{3.11}$$

$$= \sum_{Q \in \mathscr{Q}} \prod_{t=1}^{T} p(\mathbf{x}_t|q_t) P(q_t|q_{t-1}) , \tag{3.12}$$

$$\approx \max_{Q \in \mathscr{Q}} \prod_{t=1}^{T} p(\mathbf{x}_t|q_t) P(q_t|q_{t-1}) , \tag{3.13}$$

where $Q = (q_1, \cdots, q_t, \cdots, q_T)$ denotes a sequence of HMM states. Equation (3.12) is obtained by making i.i.d. and first order Markov assumptions. Equation (3.13) is obtained by applying the Viterbi approximation. As depicted in Figure 3.3, in conventional HMM, the states, i.e. the hidden variables, are discrete, while the observations are continuous modeled by Gaussian distribution for the emission scores $p(\mathbf{x}_t|q_t)$; $P(q_t|q_{t-1})$ are the transition probabilities.

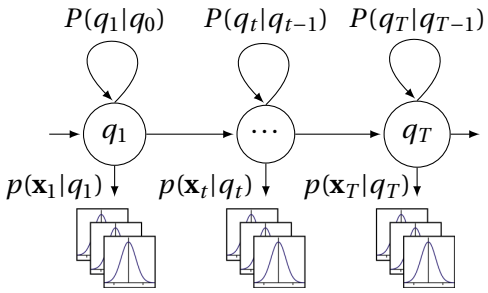

Figure 3.3 – Illustration of the hidden Markov model with his emission scores, $p(\mathbf{x}_t|q_t)$, and his transition probabilities, $P(q_t|q_{t-1})$.

In the speech recognition work [93], it was elucidated that estimation of $p(\mathbf{x}_t|q_t)$ can be interpreted as obtaining a match between $\mathbf{x}_t$ and q_t in a latent symbol set (called *acoustic unit* set $\mathscr{A}$

in [93]). More precisely, estimation of $p(\mathbf{x}_t|q_t)$ can be factored into two parts as follows

$$p(\mathbf{x}_t|q_t) = \sum_{d=1}^{D} p(\mathbf{x}_t, a^d|q_t) \ , \tag{3.14}$$

$$= \sum_{d=1}^{D} p(\mathbf{x}_t|a^d)P(a^d|q_t) \ , \tag{3.15}$$

where $a^d \in \mathscr{A}$ and $\mathbf{x}_t$ is supposedly independent of $q_t|a^d$. Intuitively, a^d are clustered context-dependent phones in a context-dependent phone-based speech recognition system [93].

As elucidated later in Chapter 5, in sign language processing, the latent symbol space tends to model the sign language production space, while the HMM states tend to model sign language perception space.

3.2.1 Estimation of $p(\mathbf{x}_t|a^d)$

The local emission score $p(\mathbf{x}_t|a^d)$ can be estimated using different techniques. In this thesis, we exploit using Gaussian Mixture Models (GMM) and ANN. As mentioned above, the conventional HMM uses Gaussian distribution as output probability, i.e.

$$p(\mathbf{x}_t|a^d) = \sum_{n=1}^{N} c_n^d \mathcal{N}(\mathbf{x}_t; \mu_n^d, \Sigma_n^d) \ , \tag{3.16}$$

where N denotes the number of Gaussian components per mixture for each state; c_n, μ_n and Σ_n denote respectively the mixture weight, mean and covariance of the n^{th} Gaussian. This approach is referred as Hidden Markov Model / Gaussian Mixture Models (HMM/GMM) approach [92].

In the hybrid Hidden Markov Model / Artificial Neural Network (HMM/ANN) approach [14, 62, 60, 123], an ANN is used to estimate the posterior probabilities $\mathbf{z}_t = [P(a^1|\mathbf{x}_t) \cdots P(a^d|\mathbf{x}_t) \cdots P(a^D|\mathbf{x}_t)]$ which are then converted to scaled-likelihoods (sl) of HMM states, i.e.,

$$fp_{sl}(\mathbf{x}_t|a^d) = \frac{P(\mathbf{x}_t|a^d)}{p(\mathbf{x}_t)} = \frac{P(a^d|\mathbf{x}_t)}{P(a^d)} \ ; \tag{3.17}$$

3.2.2 Estimation of $P(a^d|q_t)$

The relation between the production units and the HMM state, $P(a^d|q_t)$, can be defined by a deterministic map. It is the case in the conventional HMM-based approaches (HMM/GMM, hybrid HMM/ANN) where $P(a^d|q_t)$ is defined as prior knowledge by the Kronecker delta

distribution, i.e.

$$P(a^d|q_t = i) = \begin{cases} 1, & \text{if } d = i \ ; \\ 0, & \text{otherwise} \ ; \end{cases} \tag{3.18}$$

where i represents a hidden state.

Training
Given the SLR HMM-based equation (3.13), the Expectation-Maximization (EM) is used to learn the HMM parameters in all approaches; more precisely the Viterbi EM algorithm with log-likelihood as cost function $L = \log(p(X|S))$. All the HMM-based approaches were implemented using HTK [130].

Decoding
The most probable sign or sign sequence S^* is obtained by finding the most probable state sequence Q, i.e.,

$$S^* = \underset{S \in \mathscr{S}}{\arg\max}\, p(X|S)P(S) \ , \tag{3.19}$$

$$\approx \underset{Q \in \mathscr{Q}}{\arg\max} \prod_{t=1}^{T} p(\mathbf{x}_t|q_t)P(q_t|q_{t-1}) \ , \tag{3.20}$$

$$\approx \underset{Q \in \mathscr{Q}}{\arg\max} \sum_{t=1}^{T} \Big(\log(p(\mathbf{x}_t|q_t)) + \log(P(q_t|q_{t-1})) \Big) \ . \tag{3.21}$$

To do so, the Viterbi algorithm is applied in all approaches.

3.3 Posterior Feature-based Sign Language Recognition Framework

The posterior feature-based approach deals with modeling of $\mathbf{z}_t = [P(a^1|\mathbf{x}_t) \cdots P(a^d|\mathbf{x}_t) \cdots P(a^D|\mathbf{x}_t)]$, i.e. posterior probability estimates of latent symbols. The are two different ways to model them: (a) tandem feature-based approach and (b) Kullback-Leibler divergence-based Hidden Markov Model (KL-HMM) approach.

3.3.1 Tandem feature-based approach

In the tandem feature-based approach, first the latent symbols a^d are classified leading to the posterior probabilities $\mathbf{z}_t = [P(a^1|\mathbf{x}_t) \cdots P(a^d|\mathbf{x}_t) \cdots P(a^D|\mathbf{x}_t)]$. Then the posterior features are Gaussianized and decorrelated as follows,

$$\mathbf{x}_t^{\text{tandem}} = \text{KLT}(\log(\mathbf{z}_t)) \ , \tag{3.22}$$

where KLT denotes Kahunen Loeve Transform [47]. The decorrelation step also optionally reduces the feature space dimension. These features then serve as features observations for a HMM/GMM system. Although the tandem feature-based approach originally emerged from speech recognition studies, the tandem feature-based approach has been applied for handshape information-based continuous sign language recognition [62].

3.3.2 KL-HMM-based approach

In the KL-HMM approach, the latent symbol probability vector $\mathbf{z}_t = [P(a^1|\mathbf{x}_t) \cdots P(a^d|\mathbf{x}_t) \cdots P(a^D|\mathbf{x}_t)]$ is directly modeled by an HMM [4, 3], where each HMM state i is parameterized by a categorical distribution $\mathbf{y}_i = [y_i^1 \cdots y_i^d \cdots y_i^D]$. The local score is based on Kullback-Leibler (KL)-divergence:

$$S_{KL}(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} y_i^d \log(\frac{y_i^d}{z_t^d}) \ . \tag{3.23}$$

KL-divergence being an asymmetric measure, there are also other ways to estimate the local score [3]:

1. Reverse KL-divergence (RKL):

$$S_{RKL}(\mathbf{z}_t, \mathbf{y}_i) = \sum_{d=1}^{D} z_t^d \log(\frac{z_t^d}{y_i^d}) \ ; \tag{3.24}$$

2. Symmetric KL-divergence (SKL):

$$S_{SKL}(\mathbf{y}_i, \mathbf{z}_t) = \frac{1}{2} \ (\ S_{KL}(\mathbf{y}_i, \mathbf{z}_t) + S_{RKL}(\mathbf{z}_t, \mathbf{y}_i) \ ) \ . \tag{3.25}$$

Training and Decoding
The parameters of KL-HMM are estimated with the Viterbi EM algorithm by minimizing a local score based on KL-divergence. The decoding is performed using standard Viterbi decoder using the KL-divergence based local score. For more details, the reader is referred to [4, 3]. All the KL-HMM-based approaches were implemented using an in-house modified version of HTK.

As elucidated in [93], the categorical distributions of the HMM states capture a probabilistic relationship between the HMM state and the latent symbols, i.e., $\mathbf{y}_i = [y_i^1 \cdots y_i^d \cdots y_i^D] = [P(a^1|q_t = i) \cdots P(a^d|q_t = i) \cdots P(a^D|q_t = i)]$.

3.4 Summary

In this chapter, we first presented extraction of manual features. More precisely, extraction of position and velocity features from the skeleton information to model hand movement information and extraction of handshape subunits posterior probabilities to model handshape information. We then presented different hidden Markov model-based approaches used in this thesis for sign language processing. One of the challenges in using HMM for sign language processing is the lack of prior knowledge to preset HMM topology. The following chapter addresses this challenge.

4 Data-driven HMM Topology

RQ: How to determine the appropriate HMM topology for modeling the hand movement information of signs and gestures?

HMM offer a natural solution for SLR with their power in handling sequential and multimodal data. They are extensively used and have proven successful in the SLR domain [80, 28]. One of the challenges of using HMM in sign language processing is that sign languages are inherently under-resourced i.e. few well developed resources with several signers are available, and HMM require a certain amount of training data for robust parameter estimation. Another challenge is to select the structure of the HMM, i.e. the number of states, which directly can affect the performance of SLR system. Unlike speech processing, where the spoken words are represented as a sequence of subword units (e.g. phones) and the subword units are modeled through an HMM with minimum duration constraint [14], there is no such prior knowledge for sign language. As discussed in detail in Section 4.1, in many studies the number of states in the HMM is fixed for all the signs in the dataset. This may not be optimal, as the temporal structure of signs can differ, akin to temporal differences in spoken words.

This chapter focuses on addressing the challenge related to defining or determining the HMM topology. Specifically, we develop an HMM-based approach where, during the training phase, each sign is modeled by a set of HMM with different number of states. During the recognition phase, the SLR system determines the number of states for each sign independently such that the joint likelihood of the HMM state sequence and the feature observation is maximized. In other words, the approach selects the best matching HMM during testing time. The motivation being that, as there is no prior knowledge to determine the HMM topology, is to treat the number of states for each sign as a hidden information. Further, for a single sign (or lexical entity) there can be signer variations. For instance, signers can sign at different speeds (fast or slow) while varying

the hand movement. Having multiple HMM per sign could also potentially handle signer variation. To draw an analogy to spoken language processing, speech recognition systems typically handle pronunciation variation (introduced by speakers) by having multiple pronunciations as well as by changing minimum duration constraints [108]. Besides that, we also propose incorporation of a transition model, similar to silence modeling in speech recognition [130], to model portions of visual signal before and after production of signs. We validate the proposed approach through hand movement feature-based sign language/gesture recognition studies on three different databases.

The chapter is organized as follows: in Section 4.1, we present the related work on SLR and HMM modeling. Our proposed approach is explained in Section 4.2. Section 4.3 and Section 4.4 present the experimental setup and the results and analysis of the experiments, respectively. Finally, Section 4.5 summarizes the key findings. The material presented in this chapter is based on the following publication:

> *An HMM approach with inherent model selection for sign language and gesture recognition*, Sandrine Tornay, Oya Aran and Mathew Magimai.-Doss, in: Proceedings of the International Conference on Language Resources and Evaluation LREC 2020, 2020

4.1 Related Work

In most of the works on SLR that use HMM, a left-to-right HMM structure with or without skip states has been used. The number of states of the HMM has generally been fixed for all the signs/subunits in the dataset. In [111], the authors compare three different HMM topologies with different number of states, without presenting any model selection approach: fully connected, left-to-right with skip states and left-to-right without skip states and conclude that the left-to-right HMM provides the best performance, confirming the popularity of the left-to-right HMM for SLR. Only a couple of works in the literature have investigated a model selection approach for HMM for SLR. In [101], the authors present a state splitting algorithm for HMM. In their experiments on a dataset of signs from Australian Sign Language, their proposed approach is faster and achieves better performance than the conventional HMM Baum-Welch training. In [74], the HMM topology is automatically constructed from an initial topology by modifying it using segments, which are formed based on the segmentation of hand motion. In [120], low rank approximation is used to determine the key frames of a sign which guides the selection of the number of states of HMM independently for each sign. In [68], an entropy-based k-means algorithm is used to determine the HMM number of states. With this approach, each sign is represented by an HMM with different number of states. Additionally, an artificial bee colony algorithm is used together with the Baum-Welch algorithm to determine the HMM structure. Their experiments show that the proposed approach achieves better performance than a left-to-

right HMM structure with fixed number of states. However, from the experiments reported, it is no clear how much of the performance increase comes from the selection of the number of states and from the determination of the HMM structure through the swarm optimization algorithm.

4.2 Proposed Approach

The proposed HMM-based model selection approach assumes that each sign could have different number of stationary states. Intuitively, when considering hand movement information modeling, an HMM state can represent a specific position, orientation or location of the hands depending on the input observation features. Therefore, we make the assumption that the complexity of a sign influences the appropriate number of states used to model it. To the authors' knowledge, no method in the literature today allows to set this number beforehand. An exhaustive search using cross validation is not feasible in the sign language domain as the number of signs in the datasets is typically high. Thus, instead of setting this number beforehand, the proposed model selection approach selects the appropriate one in an interval of possibilities at the recognition stage. More precisely, an interval of possible number of states is first chosen, let's say N_{min} to N_{max}. Then, $\forall n \in [N_{min}, N_{max}]$, a left-to-right HMM with n states is trained for each sign. Then at the recognition stage, the model yielding maximum likelihood is chosen as the recognized output (see Figure 4.1). Thus $S \cdot (N_{max} - N_{min} + 1)$ models are tested, in comparison to S in the first approach, where S is the number of unique signs in the dataset.

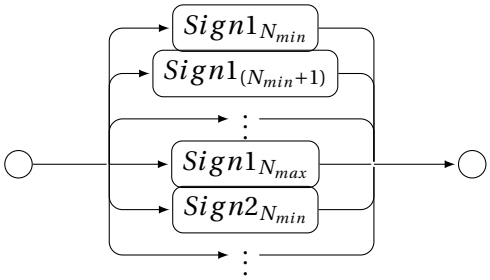

Figure 4.1 – Recognition network of the proposed model selection approach.

Besides the choice of the number of states, the quality of the data segmentation can also affect the sign-based model. Indeed, the exact start and end of a sign is not perfectly defined, especially in a continuous context where each sign is being followed by other signs. In the isolated context, the segmentation is not necessarily optimized leading to the same problem. Thus, in both cases, there is a transition phase at the beginning and at the end of the performed sign. This period can represent the absence of movement/hand gesture or even some slight insignificant movement. To handle this issue, we propose to add a transition model, common to each sign, before and after each sign-based HMM. For preserving the continuity of the entire model, we modeled it as a

three-state left-to-right HMM with one-state-skip (see Figure 4.2 for the structure). This structure is inspired from speech processing to handle short pauses and silent modeling [130].

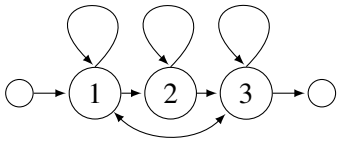

Figure 4.2 – HMM topology of the transition model.

4.3 Experimental Setup

As a first step, we conducted hand movement information-based isolated sign language recognition and gesture recognition tasks on three databases to validate the proposed approach. Chapters 6-7 further validate this approach in the context of modeling handshape information.

4.3.1 Datasets

We used the Chalearn14 gesture database and the DGS and HospiSign sign language databases, described earlier in Section 2.3. The partition of the training and testing data is given in Table 4.1. In DGS and HospiSign studies, we conducted leave-one-signer-out cross validation experiments; therefore, Table 4.1 reports the mean of the data samples used in each experiment. For the Chalearn14 case, we kept the given data partition of the challenge.

Table 4.1 – Partition of the Chalearn14, DGS and HospiSign databases into training and testing data samples

	Train	Test
Chalearn14	9306	3579
DGS	2586	227
HospiSign	1049	210

4.3.2 Hand movement feature extraction

For extracting the hand movement features, as described earlier in Section 3.1.1, we rely on the tracked 3D coordinates of a human skeleton. The 3D trajectories of the two hands as well as the other skeleton joints such as head, neck, shoulders and hips form the basis for our continuous features of hand motion information, in particular hand position and velocity. The hand movement

feature extraction process yields 36-dimensional feature vector at each time frame.

4.3.3 Systems

We studied three different systems, where the method used to infer the number of states is different:

- The *msHMM* system stands for the proposed model selection approach, where the inference of the number of states N is different for each sign and is not fixed beforehand. Only the interval of possibilities from N_{min} to N_{max} has to be defined.
- The *sdHMM* system stands for standard HMM topology where the number of states N is the same for all the sign and is fixed beforehand. $\forall n \in [N_{min}, N_{max}]$, a n states HMM is trained and the one that yields the best performance on the test set serves as a baseline system.
- The *kmHMM* system stands for the approach developed in [68] that we implemented to further validate the proposed selection method. In this approach, the number of states N is set using the entropy-based k-means algorithm. For fair comparison, the k was selected in the same range, i.e. $k \in [N_{min}, N_{max}]$.

In all the above presented systems: HMM refers to left-to-right HMM. The range of the number of states was $N_{min} = 3$ to $N_{max} = 13$. In all cases, each HMM state emission distribution is modeled with a single multivariate Gaussian with diagonal covariance matrix. The performance of the systems was evaluated in terms of recognition accuracy (presented in Section 2.4).

We also investigated a variant of the above described systems, namely, *tr-msHMM*, *tr-sdHMM* and *tr-kmHMM*, that incorporate transition models at the beginning and end of each sign-based model. Since by adding the three states transition model we increase the number of states of the sign-based model by at least four states (two states before and after each model, see Figure 4.2), we decided to adapt the range of possibilities, leading to $N_{min} = 3$ to $N_{max} = 9$. Figure 4.3 depicts the recognition network of the *tr-sdHMM* and *tr-kmHMM* systems, while Figure 4.4 the *tr-msHMM* system.

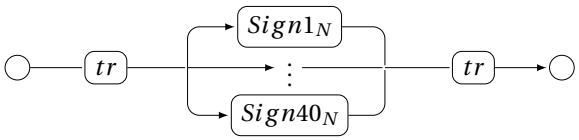

Figure 4.3 – Recognition network of the *tr-sdHMM* and the *tr-kmHMM* system.

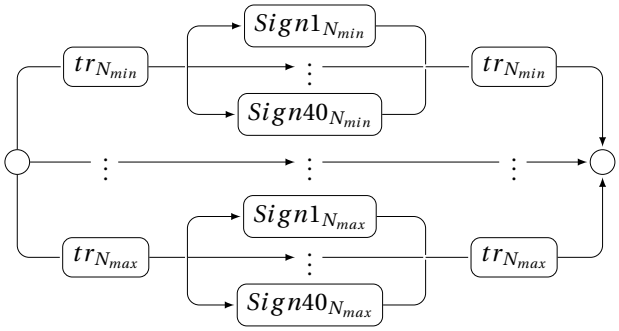

Figure 4.4 – Recognition network of the *tr-msHMM* system.

4.4 Results and Analysis

In this section, we first present the recognition accuracies of the different systems on the Chalearn14, DGS and HospiSign databases. Next, we contrast the performances obtained by our approach with the existing studies reported on those databases to demonstrate that the results obtained by our systems are competitive.

4.4.1 Comparison of systems

Figures 4.5 presents the recognition accuracy of the three systems without the transition model, all are signer-independent. Firstly, looking at the performances of the *sdHMM* system, we can deduce that the number of states has an impact on the recognition accuracy and tends to saturate. Secondly, we can observe that the proposed approach (*msHMM*) consistently outperforms the approach of setting number of states based on k-means (*kmHMM*). Furthermore, we can also observe that the *msHMM* system yields performance comparable to the *sdHMM* system with fixed number of states yielding the best performance on the test data.

Figures 4.6 presents the recognition accuracy of all the systems containing the transition model. We can observe that the recognition performance of all the systems considerably improve. As the transition model is common to all signs, the improvement can be attributed to the modeling of sign-independent irrelevant information at the beginning and end of the visual signal. When comparing *tr-msHMM*, *tr-kmHMM* and *tr-sdHMM* systems, the trend remains the same, i.e. *tr-msHMM* system is better than *tr-kmHMM* system and is comparable to the *tr-sdHMM* system.

For the sake of completeness, Table 4.2 summarizes the recognition accuracy with standard deviation for all the systems.

Figure 4.7 shows the histogram of the number of states of the HMM selected during recognition phase of *tr-msHMM* system for Chalearn14, DGS and HospiSign. As expected, it can be observed

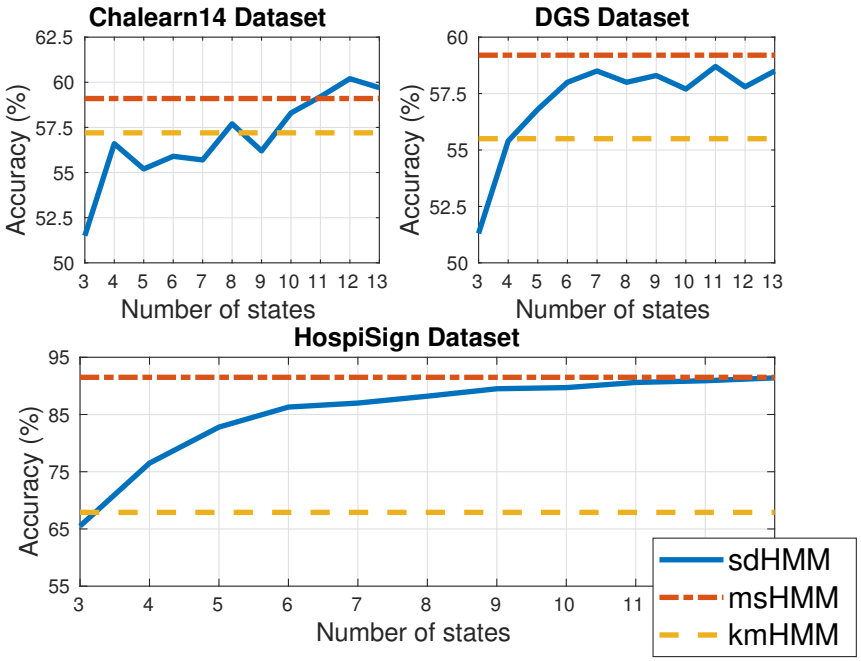

Figure 4.5 – Recognition accuracy of the *sdHMM*, the *msHMM* and the *kmHMM* systems.

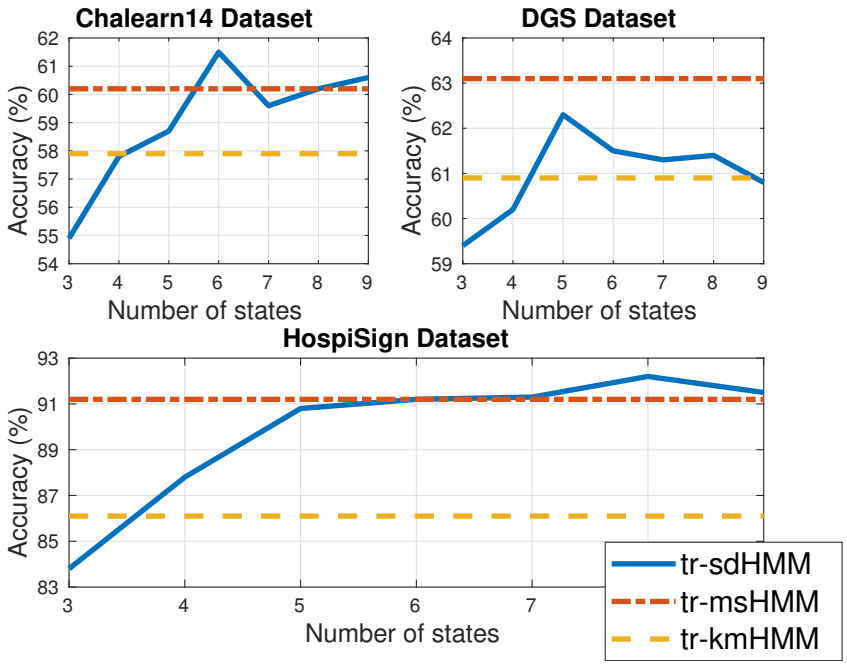

Figure 4.6 – Recognition accuracy of the *tr-sdHMM*, the *tr-msHMM* and the *tr-kmHMM* systems.

Table 4.2 – Recognition accuracy of the systems on the Chalearn14, DGS and HospiSign dataset.

	Chalearn14	*DGS*	*HospiSign*
msHMM	59.1	59.2 ± 9.6	91.5 ± 7.0
sdHMM (# states)	60.2 *(12)*	58.7 ± 11.5 *(11)*	91.4 ± 6.0 *(13)*
kmHMM	57.2	55.5 ± 10.5	67.9 ± 4.4
tr-msHMM	60.2	63.1 ± 10.3	91.2 ± 6.5
tr-sdHMM (# states)	61.5 *(6)*	62.3 ± 9.8 *(5)*	92.2 ± 4.9 *(8)*
tr-kmHMM	57.9	60.9 ± 9.5	86.1 ± 6.3

that HMM with different number of states are selected at run time. In the case of DGS and HospiSign models, the histogram is skewed towards higher number of states. However, it is not the case for Chalearn14. One possible reason for that could be that Chalearn14 has simple gestures (hand up and down movement) in an uncontrolled environment.

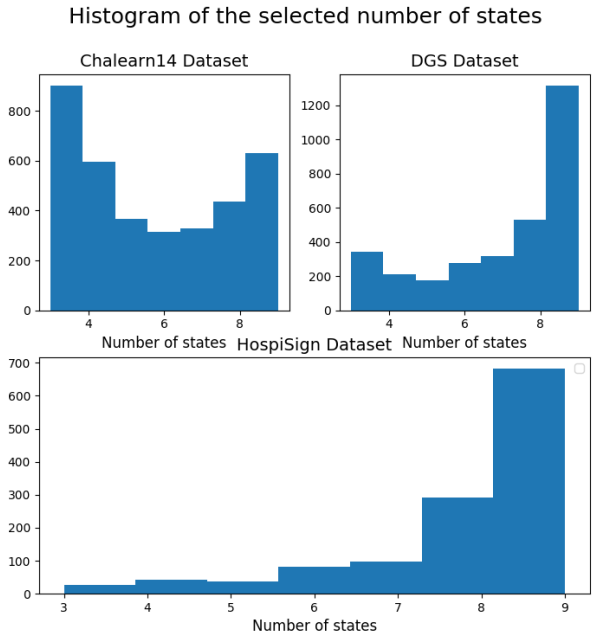

Figure 4.7 – Histogram of the selected number of states during the recognition process using the *tr-msHMM* systems.

4.4.2 Comparison to existing studies

In this section, we contrast the performances obtained on DGS dataset and HospiSign dataset to existing studies reported on these datasets. These studies have used the same protocols as we have. In the case of Chalearn14, the evaluation is based on Jaccard index that involves joint

evaluation of segmentation and recognition of gestures [38]. A fair comparison is not feasible, as it is difficult to separate the contribution of segmentation errors and recognition errors of the systems reported in [38]. Thus, we do not contrast for Chalearn14.

DGS Database

Table 4.3 compares performance with our models with two other works on the DGS dataset. In [79], the authors used a multi-class sequential pattern tree with boosting for classifying signs, using binary features based on the hand motion and location information. In [29], the authors proposed a subunit-based approach using a sequential pattern boosting classifier, where the subunits are extracted based on the different modalities that make up a sign, i.e. handshape, hand location, hand motion, and hand arrangement. It is important to note that the dataset used in [29] contains signs from one extra signer, which we do not have access to in our dataset. Based on the reported signer independent performance of 49.4% in [29], we calculated the accuracy range for the remaining 14 users (assuming that the accuracy on the 15^{th} user could take a value between 0% and 100%). This calculation gives us a range of $[45.7, 52.9]$, which is still lower than the performance achieved by the proposed *tr-msHMM* system.

Table 4.3 – Comparison of our systems with existing studies for the DGS dataset

Method	Signer Indep. (%)
Sequential Pattern Trees [79]	55.4 ± 8.1
Boosted Subunits [29]	49.4 ± 8.5
tr-msHMM system	63.1 ± 10.3

HospiSign Database

Table 4.4 compares the performance of our system with the performance reported in [18]. Briefly, in [18], various manual features such as handshape, hand position and hand movement were extracted and temporal modeling using either DTW or temporal templates was performed. In the case of using DTW, the signs were classified using k-Nearest Neighbours (k-NN). We contrast to the system where only hand movement information is modeled. We also trained a *tr-msHMM* system that uses the same hand movement and hand joint feature as in [18]. In both cases, we can observe that the proposed *tr-msHMM* system yields performance close to the best reported system using only hand movement information.

Table 4.4 – Comparison of our systems with existing studies for the HospiSign dataset

Method	Signer Indep. (%)
Hand Joint and Movement Distances [18]	93.8 ± 6.4
tr-msHMM system	91.2 ± 6.5
tr-msHMM system using the same "Hand Joint and Movement Distances" features as [18]	91.6 ± 6.1

4.5 Summary

This chapter presented a HMM-based model selection approach where, during training, each sign is modeled by a set of HMM with different number of states and the best matching model is automatically selected during the recognition phase based on maximum likelihood criteria. We also investigated the use of a transition model taking inspiration from silence modeling in speech processing. Our investigations on sign language recognition and gesture recognition tasks on three different databases showed that the proposed model selection approach yields better systems than the approach of presetting the number of HMM states using k-means and yields systems competitive to the baseline system with fixed number of states determined on the test set. Furthermore, incorporation of a transition model to model portion of visual signal before and after the production of each sign helps in improving the performance of systems. So, in the remainder of the thesis work, we used the left-to-right HMM topology with transition model.

It is worth mentioning that, although the investigations were carried out on isolated signs, gestures and phrases, the approach can be extended to continuous sign language processing. The different HMM for each sign can serve a similar role as multiple pronunciations for each word in continuous speech recognition systems. Thus, the model selection aspect in continuous sign language processing can be handled by the decoder in the same manner as selecting the pronunciation for a word in a continuous speech recognition system. In the following chapter, we will demonstrate that the model selection approach can potentially be exploited for hand movement subunits extraction.

5 Hand Movement Subunits Derivation

RQ: How to discretize and model hand movement information as subunits?

As described in Chapter 3, the movement of both hands is a relevant structure of the sign production that needs to be modeled. In the literature, it is well understood that the handshape information can be modeled as a sequence of subunits [62, 20] such as HamNoSys [45, 59]. However, the continuous aspect of the movement makes modeling of movement information as subunits difficult. Indeed, sequence modeling for signs and gestures is an open research problem. In that direction, there is a sustained effort towards modeling signs and gestures as a sequence of subunits. In order to develop efficient sign language processing systems, it is desirable to model signs as a sequence of subunits, akin to phoneme- or phone-based speech processing [11]. As, subunits allow robust parameter estimation. It could also remove the constraint that all signs in the lexicon needs to be observed during training. Furthermore, subunits can allow data sharing across languages [62]. The multistream nature of sign language implies that the development of such a subunit set is a highly challenging task.

In this chapter, we present a novel HMM-based approach to extract hand movement data-driven subunits from skeleton information by building upon the inherent ability of HMM to segment time series into stationary segments for sign modeling. In this approach, no prior knowledge of the number of subunits or segmentation or linguistic annotation is used. Rather, only pairwise comparison between signs production, i.e. whether two productions correspond to the same sign or not is used. The approach involves: (a) extraction of position and motion features from 3D skeleton information; (b) inferring a left-to-right HMM for each sign by modeling the position and motion features; (c) clustering the states of the HMM across the signs through a measure of discrimination to infer subunits and representing each sign in terms of those subunits; and finally

(d) visualization based on HMM-based synthesis. We developed the proposed framework for subunits extraction for both sign language processing and speech processing. Specifically, in the sign language study the above-mentioned recognition-synthesis framework for hand movement subunits extraction and analysis was developed. Whilst, in the spoken language study we demonstrated that the framework can lead up to phone set discovery and pronunciation lexicon development. For the sake of clarity, the investigations on spoken languages are presented in Appendix A.

The remainder of the chapter is organized as follows: Section 5.1 presents the related work. Section 5.2 presents the proposed framework where in the inference of the left-to-right HMM step the number of states is fixed based on the recognition accuracy on the training and development data. Section 5.3 presents the experimental setup and the results on sign language recognition task. In Section 5.4, we investigate if subunits exhibit language independent property by conducting a cross-lingual sign language recognition task. In Section 5.5, we investigate the potential of using the model selection approach presented in Chapter 4 in the inference of the left-to-right HMM step. Finally, in Section 5.6 we discuss the salient findings and conclude. This chapter is based on the following publications:

Subunits inference and lexicon development based on pairwise comparison of utterances and signs, Sandrine Tornay and Mathew Magimai.-Doss, in: Information, Special Issue: Computational Linguistics for Low-Resource Languages, 10:298, 2019

Data-driven movement subunit extraction from skeleton information for modeling signs and gestures, Sandrine Tornay, Marzieh Razavi and Mathew Magimai.-Doss, Research Report, Idiap-RR-02-2019

5.1 Related Work

The focus of this chapter lies in automatic derivation of hand movement subunits for sign language processing. In the literature, there are two strands of research in that direction.

The first strand of research makes the assumption that some annotation of signs is available. Pitsikalis et al. [90] incorporated phonetic transcription into data-driven subunits. They first converted Hambourg Notation System (HamNoSys) symbols into Posture-Detention-Transition-Steady Shift model. Then they combined these structured sequences of labels with visual tracking features for timing information via an HMM-based system to obtain the phonetic subunits. Cooper et al. [29] used hand labeled data and compared three types of subunits: appearance-based, 2D tracking-based and 3D-tracking based. Two sign-level classifiers were tested: an HMM-based

approach and the sequential pattern boosting. Koller et al. [61] used gloss annotations and gloss time boundaries to generate sequences of subunits using HMM-based modeling and expectation-maximization algorithm. Elakkiya and Selvamani [36] extracted manual and non-manual features by using parallel HMM and introduced a novel Bayesian parallel HMM to combine the visual and linguistic transcriptions of the sign lexicon to form a subunit gesture base.

The second strand of research involves extraction of subunits without using annotation information. In this case, subunits extraction typically involves unsupervised segmentation and clustering. There exist two lines of thoughts based on the order in which segmentation and clustering steps are carried out, i.e.,

- *Clustering followed by segmentation:* Bauer and Kraiss [8] used *k*-means algorithm to cluster the data where each cluster is then represented as a fenonic baseform. Temporal structure is then achieved with the HMM-based structure defined based on this fenonoic model [51]. Han et al. [53, 7, 44] used hand motion speed and trajectory to locate subunit boundaries and then temporal clustering by DTW is adopted to merge similar subunits.

- *Segmentation followed by clustering:* Sako and Kitamura [98] extracted different subunits by training a multi-stream isolated sign HMM for each word where the feature vector of each frame is split into three phonetic stream, and by clustering each state of the multistream using an inter-state distance with a tree-based algorithm in order to tie the states. Fang et al. [42] segmented signs using HMM in which each state represents one segment. Then they used a temporal clustering algorithm based on modified *k*-means algorithm where DTW is employed as the distance computation criterion. In that study, CyberGloves and Pohelmus 3SPACE-position trackers were used. Based only on simple position measurements obtained from the video, Theodorakis et al. [111] used, as an initial segmentation step, the model-based segmentation proposed in [42], and then employed a hierarchical clustering of whole dynamic models (HMM) to find the shared segments.

5.2 Proposed Subunit-based Lexicon Development

This section presents the proposed computational methodology for hand movement subunits extraction and analysis. The proposed methodology formulates the subunits extraction problem as follows: Given only a set of sign productions and the knowledge that any two pair of sign productions correspond to the same lexeme or not, how to derive the set of subunits and model each distinct sign as a sequence of subunits? The proposed methodology consists of four steps:

Step 1: first, a sequence of feature vectors is extracted for each sign production. The feature vectors for hand movement are based on the 3D skeletal information (see Chapter 3.1.1);

Step 2: given the sequence of feature vectors for each sign production, a HMM is obtained for each distinct sign in the set. This step exploits the idea that HMM inherently segment a time series into stationary segments and sign recognition can be performed with word level HMM;

Step 3: next, the states are clustered into subunits by pairwise comparison and a sequence model in terms of clustered subunits is obtained for each distinct sign; and finally

Step 4: visualization based on HMM-based synthesis is employed.

As illustrated in Figure 5.1, Step 2 and 3 can be grouped together and seen as a step of deriving automatic subunits and development of an automatic subunit-based lexicon. More precisely, in Step 2 sign-level HMM for each sign are determined. This is done by modeling each state by a single Gaussian distribution with diagonal covariance and finding the number of states, n, such that the recognition accuracy saturates on the training and the development data. This process yields the same number of states n for all the distinct signs. The motivation of selecting this approach comes from the saturation observation of the recognition accuracy of the standard HMM system (*sdHMM*) in Chapter 4. As mentioned earlier, the model selection proposed in Chapter 4 can be applied to set the HMM topology in Step 2, this investigation is proposed in Section 5.5.

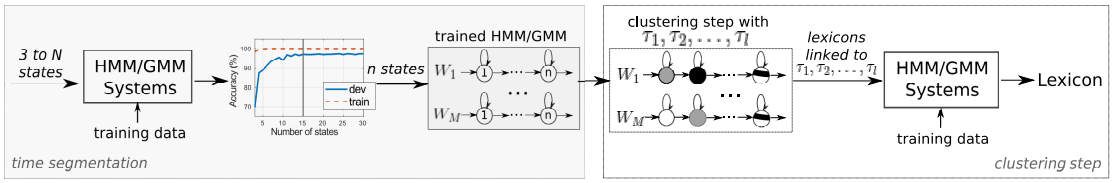

Figure 5.1 – Illustration of the subunit-based lexicon generation.

Given all the single Gaussians of the signs HMM states, Step 3 clusters them through a measure of discrimination. More precisely, this is done by computing, between each pair of Gaussian distributions, the Bhattacharyya distance [9]:

$$Bhatt(\mathcal{N}_1,\mathcal{N}_2) = \frac{1}{8}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+\frac{1}{2}\ln\left(\frac{\det\boldsymbol{\Sigma}}{\sqrt{\det\boldsymbol{\Sigma}_1\det\boldsymbol{\Sigma}_2}}\right), \tag{5.1}$$

where $\mathcal{N}_1 := \mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $\mathcal{N}_2 := \mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ are two Gaussian distributions and $\boldsymbol{\Sigma} := \frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}$. The level of similarity between two HMM states is defined by a threshold τ, i.e. two HMM states are similar if the Bhattacharyya distance between the Gaussian distributions corresponding to the two states is below the threshold τ. The intuitive explanation is that two segments are modeling similar information or subunit if the probability density functions of those states are similar. This clustering step yields a set of automatic subunits and an automatic subunit-based lexicon based on τ. The hyper-parameter τ is determined in a cross-validation manner where,

1. first multiple automatic subunit-based lexicons corresponding to different values of τ are obtained;
2. a recognition system is then trained on the training data based on each of those lexicons; and
3. the lexicon that yields best recognition accuracy on the development data is chosen.

Selection of τ in this manner ensures that minimal discrimination between signs are maintained after the clustering step.

One way to validate is to recognize the subunits. Another way to do it is to visualize the hand movement subunits which allows to ascertain the identity of them; this is the goal of Step 4. The subunit-based lexicon represents the hand movement information for each observed sign “in-parts". It is not obvious to what prior linguistic knowledge those subunits could be linked. Even the HamNoSys annotation [45], which is used to transcribe signs, transcribes the whole movement information, not the movement information in-parts. We develop a method, where the trained HMM are used as a generative model to synthesize hand movement information in the 3D feature observation space by applying a Linear-Quadratic Regulator (LQR) [87, 13]. The synthesized hand movements for the signs can then be visually compared to the actual movements produced by the signers, which subsequently could be linked to HamNoSys should such transcription be available. Figure 5.2 illustrates the proposed approach to synthesize hand movement information of signs based on the derived subunits, starting from KL-HMM. As illustrated in the figure, given a sequence of Gaussian distributions, LQR finds the minimal path which passes through the sequence of the Gaussian distribution linked to each subunit that model a particular sign. The main idea here being: are the derived subunits able to synthesize the hand movement of signs such that it corresponds well with the human sign production?

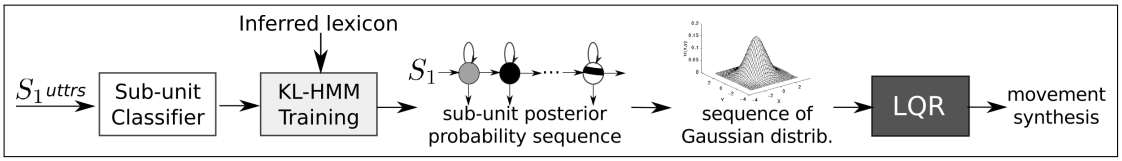

Figure 5.2 – Illustration of the hand movement synthesis approach according to the hand movement subunits.

5.3 Monolingual Study

We validated the proposed approach subunits extraction approach by conducting *signer-independent* automatic SLR study and by synthesizing and analyzing hand movement information from the resulting models on a sign language database.

5.3.1 Experimental setup

SMILE DSGS database

We evaluated the proposed approach on the SMILE Swiss German Sign Language database. More information on the database can be found in Section 2.3. As mentioned in Section 2.3, only the second pass was annotated through six categories that evaluates the acceptability of a sign according to linguistic criteria. In our experimental studies, we only used the second pass data that was annotated as Category 1 or 2, i.e. acceptable signs with the same or slightly the same form. The data was partitioned in a signer-independent manner into 1263 training set samples from 17 signers, 249 development set samples from 3 signers and 704 test set samples from 10 signers.

Systems

We built HMM/GMM [92], hybrid HMM/ANN [14] and KL-HMM systems (see Chapter 3.2 for statistical explanation of the models) to evaluate the automatic subunits based lexicon at SLR level. In each case, we built two systems:

(a) *sign level system:* using sign-level HMM states obtained in Step 2 as subunits. This system is the standard HMM (*sdHMM*) system presented in Chapter 4;

(b) *SU-based system:* using the clustered HMM states in Step 3 as subunits.

The motivation behind building sign level system is that Step 2 obtains a sign level HMM with fixed number of states n through discrimination like in Step 3, so the states of the sign level HMM can also regarded as subunits without being clustered. Such a comparison would help us to determine whether the clustering step is indeed yielding meaningful subunits or not. For all systems, we used the skeleton-based position and velocity features of both hands as input features. The resulting feature vector is of size 36, see Section 3.1.1 for details.

HMM/GMM Systems: All the HMM/GMM systems are left-to-right HMM using one mixture Gaussian distribution with diagonal covariance matrix per state as the emission distribution. In Step 2, the number of states for the sign level HMM is chosen according to the saturation of the model on the SMILE training and development data. The range of state is from 3 to 30. In Step 3, the clustering step was conducted with the hyper-parameter, τ, in the range of 0.3 to 1.3 with a 0.1 step, leading to a set of lexicons. An HMM/GMM system was trained for each lexicon and the one that yields the maximum recognition accuracy on the development set was chosen. Test set performances are reported on that lexicon.

Hybrid HMM/ANN Systems: For building the hybrid HMM/ANN systems, we first obtained the alignments in terms of the HMM states using either the sign level or the SU-based HMM/GMM systems. We then trained Multilayer Perceptron (MLP)s classifying HMM states with output non-linearity of softmax and minimum cross-entropy error criterion. We used the 36-dimensional movement features with four frames preceding context and four frames following context as the MLP input. In our experiments, we trained MLPs with different number of hidden units (600, 800, 1000) and hidden layers (0, 1, 2, 3). The number of hidden units and hidden layers as well as other hyper-parameters such as learning rate and the batch size were chosen according to the frame-level accuracy on the development set.

We estimated the scaled likelihoods in the hybrid HMM/ANN systems [14] by dividing the posterior probabilities derived from MLPs with the prior probabilities of the classes estimated from relative frequencies in the training data. These scaled likelihoods were then used as emission probabilities for HMM states during decoding.

KL-HMM Systems: The hand movement subunits posterior probabilities estimated by the MLP of hybrid HMM/ANN system are used as feature observations. The KL-HMM states represent the hand movement subunits. The cost function used to train and test was the reverse KL-divergence (see Section 3.3.2).

The evaluation measure used in this experiment is the Recognition Accuracy (RA) presented in Section 2.4.

Synthesis: We conducted visualization studies by applying LQR-based hand movement information synthesis using the *pbdlib* library, developed by Pignat and Calinon in [87] in the context of robotics.

5.3.2 Results and analysis

Table 5.1 presents the sign language RA depending on the SU-based system and sign level system on the SMILE DSGS database along with the average number of subunits in each case. It can be observed that the SU-based system with around 14% less HMM states performs comparable to sign level system for all the systems: HMM/GMM, hybrid HMM/ANN and KL-HMM. This indicates that the clustered subunits-based lexicon obtained in Step 3 maintains discrimination across signs. Furthermore, HMM/GMM systems trained using subunits posterior probability as feature observation further improves the recognition accuracy.

To get an insight into how consistent are the derived subunits, we synthesized the 3D hand position movement of the unobserved test samples with a LQR using the sequence of Gaussian distributions linked to the sequence of subunits of the sign (Step 4). The starting point for the hand

Table 5.1 – Hand movement clustered subunits-based and sign level HMM/GMM, hybrid HMM/ANN and KL-HMM systems performance in terms of recognition accuracy on the SMILE DSGS database

	Clustered SU-based system	Sign level system
HMM/GMM	51.3	49.4
Hybrid HMM/ANN	51.6	53.0
KL-HMM	55.8	57.4
Average # subunits	*1945*	*2256*

movement synthesis is the starting point of a sign's production coming from a training sample. The duration for each state is the average number of time frames estimated by aligning the states in the sign to the training samples. Two signs are presented: TAXI which is a well-recognized sign (100% of recognition) and PAPIER which is a poorly recognized sign (0% of recognition). To facilitate the visualization, we depict in Figure 5.3 the (x, y) position of the dominant hand as well as for comparison three examples of the respective sign samples (soft lines); the $z-$axis, the depth of the sign production, being not relevant for these two particular signs.

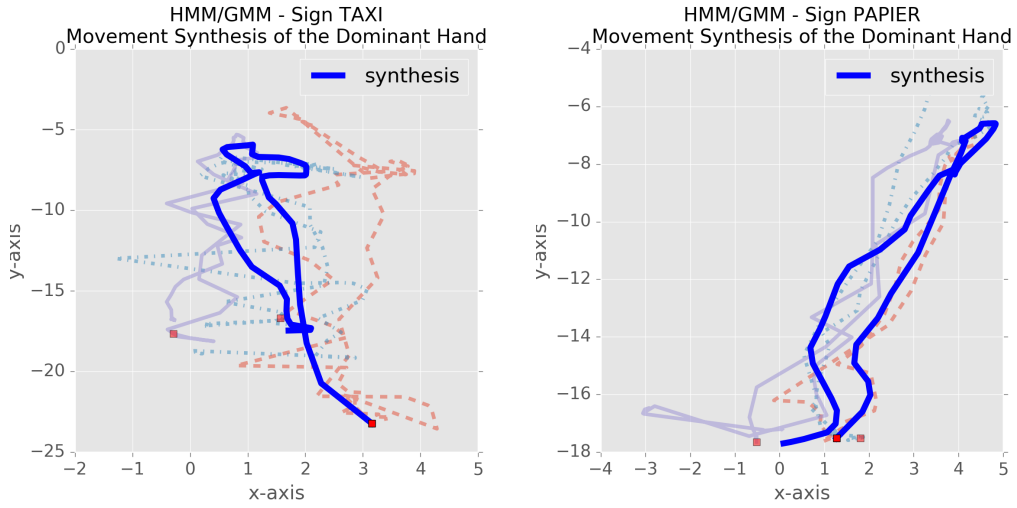

Figure 5.3 – Hand movement synthesis of the dominant hand for the well-recognized sign TAXI (left) and the poorly-recognized sign PAPIER (right) using the Gaussian distribution sequence of the SU-based HMM/GMM system. The red squares are the starting points.

As it can be seen, the hand movement of signers vary a lot.[1] Nevertheless, in both cases, the synthesized movements follow similar direction and range of movement as the hand movement

[1] For the sake of clarity, we did not show all the signers production.

of the actual signers. This suggests that the subunits are modeling the relevant hand movement information.

The sequence of the Gaussian distributions can be also obtained based on the parameters of the KL-HMM system. More precisely, the categorical distributions at each state can be used to compute a new Gaussian by using them as a weight on the subunit Gaussians. First, we selected significant components, i.e. categorical distribution components that have a probability mass greater than 0.005, and re-scaled them according to the total number of selected components, M. Then the combined mean, μ_{comb}, and diagonal covariance, σ_{comb}, are computed as:

$$\mu_{comb} = \frac{1}{M} \sum_{m=1}^{M} \mu_m \quad , \qquad \sigma_{comb} = \sum_{m=1}^{M} w_m \cdot \sigma_m;$$

where w_m are the re-scaled categorical distributions and μ_m, σ_m the mean and the diagonal covariance of the corresponding Gaussian distribution. Figure 5.4 depicts the resulting movement synthesis of the well-recognized sign, TAXI (100% of recognition) and WASCHEN a poorly-recognized sign (0% of recognition). When comparing the synthesized movement for sign TAXI across KL-HMM and HMM/GMM (Figure 5.3), the difference mainly appears at the end of the sign.

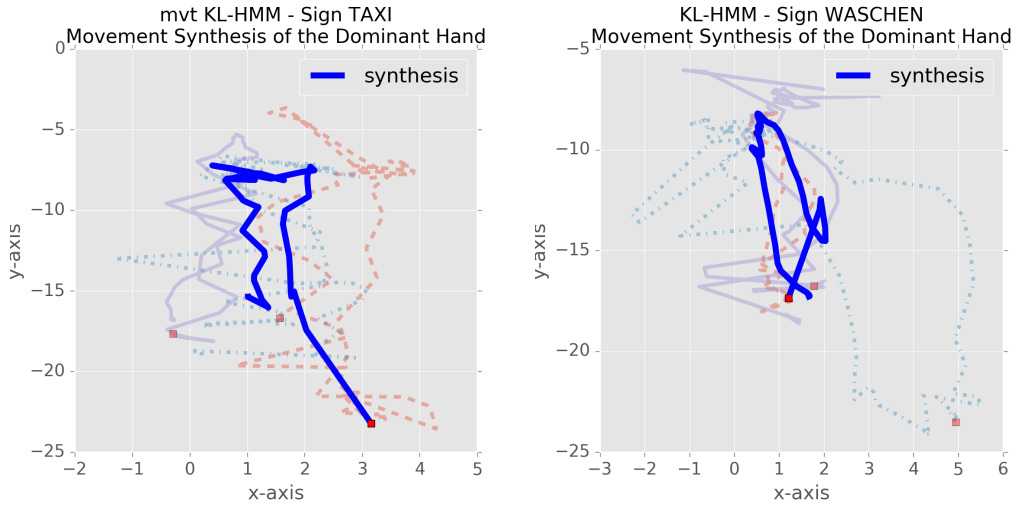

Figure 5.4 – Hand movement synthesis of the dominant hand for the well-recognized sign TAXI (left) and the poorly-recognized sign WASCHEN (right) using the Gaussian distribution sequence computed from the SU-based KL-HMM system. The red squares are the starting points.

As the synthesized movement of the dominant right hand of the poorly-recognized sign, PAPIER, is corresponding well to the movements produced by the signers, we looked at the confusion matrix to understand the reason for the poor recognition accuracy and analyzed the hand movements.

It was found that the hand movements of some signs are similar, as it can be seen in Figure 5.5.

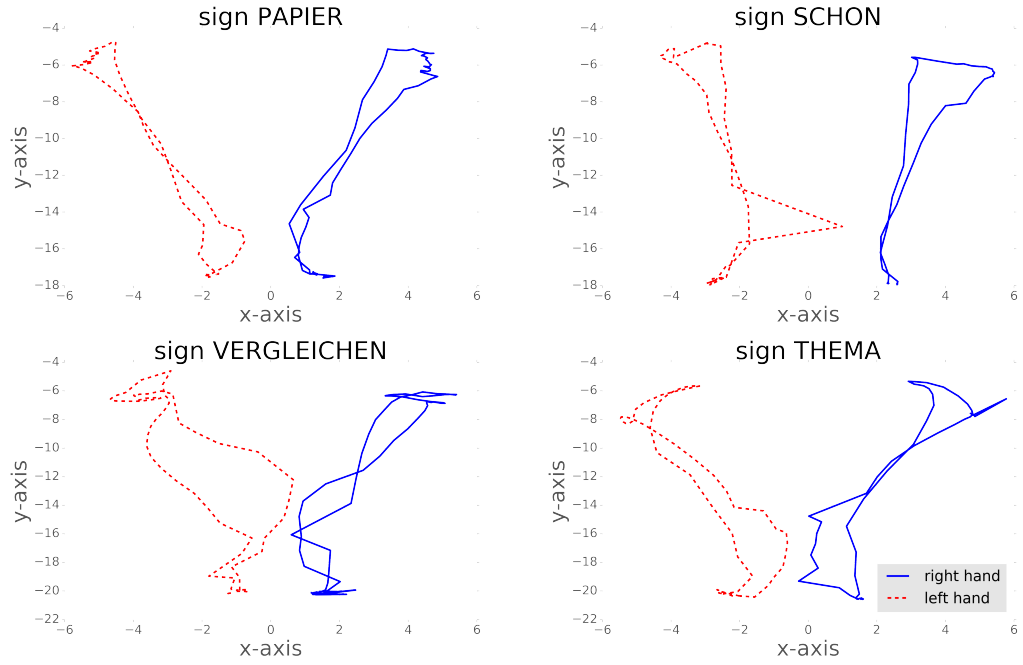

Figure 5.5 – (x, y) movement of the right and left (dashed line) hand of the signs PAPIER, SCHON, VERGLEICHEN and THEMA, respectively.

Sign language convey information through multiple channels. A single channel (e.g., only hand movement) may not be sufficient to discriminate all the signs. The confusion in terms of hand movements can be handled by adding other channel such as handshape to the KL-HMM system. This research is presented in Chapter 6.

5.4 Cross-lingual Study

Speech technologies such as, automatic speech recognition systems, benefit from the idea that subword units such as phones can be shared across languages [93]. In the spoken language part, presented in Appendix A, we have demonstrated that capability through a study using auxiliary multilingual resources. A question that arises is: whether the derived hand movement subunits exhibit similar desirable characteristics? For that we used the SMILE Swiss German sign language database and the TSL HospiSign database (see Section 2.3 for detailed databases) to perform cross-lingual sign language processing experiments.

5.4.1 Experimental setup

We developed left-to-right KL-HMM systems to investigate the ability to share the derived subunits across sign languages.

cross lingual KL-HMM Systems: The hand movement subunits are derived on TSL HospiSign database (Step 2 and Step 3); an MLP is trained on the HospiSign data to estimate TSL subunits posterior probabilities; and the states model DSGS subunits and the parameters are trained by using the TSL subunits posterior probabilities estimated on the DSGS data as feature observations. In doing so, the KL-HMM learns a probabilistic relationship between DSGS subunits and TSL subunits, and lets us to examine language-independence of derived subunits. To compensate the difference in the coordinate system recording in between both databases, a skeleton alignment is applied before the feature extraction. To do so, all the signer skeletons of both databases are aligned at the neck joint with respect to a reference Hospisign signer skeleton and then scaled by the shoulder width. We used shoulder-normalization-based features of both hands as input features. The resulting feature vector is of size 36, see Section 3.1.1 for details.

5.4.2 Results and analysis

Table 5.2 presents the results of the cross-lingual study, where the subunits are derived on the TSL HospiSign and KL-HMM system is trained on DSGS database to recognize DSGS signs with DSGS subunits lexicon. It can be observed that the performance drops considerably when compared to the monolingual case. However, the performance obtained is beyond chance level. This suggests that there exists some degree of systematic relationship between the DSGS subunits and TSL subunits but it is not sufficient to recognize well the DSGS signs. The reason for that could be: (a) differences in the coverage of hand movements across the two databases and (b) differences in recording settings. In the case of HospiSign, the signs were performed in standing position, while in the case of DSGS, the signs were performed in sitting position. Skeleton alignment may not have fully compensated for these differences.

Table 5.2 – Sign language RA on the SMILE DSGS database of the KL-HMM system trained with TSL HospiSign subunits posterior probabilities in the multilingual case and DSGS subunits in the monolingual one

	Cross-lingual system	Monolingual system
Sign RA	41.5	55.8

Figure 5.6 depicts the synthesis of the movement based on the TSL HospiSign subunits of a well-recognized sign and poorly-recognized sign, THEMA and SPIELEN respectively. In the case of sign THEMA, we can observe that the synthesized movement follow hand movement of

the actual signers. Whilst, in the case of sign SPIELEN it is not the case. One of the reasons for that could be that TSL subunits may not be covering well all the DSGS movements.

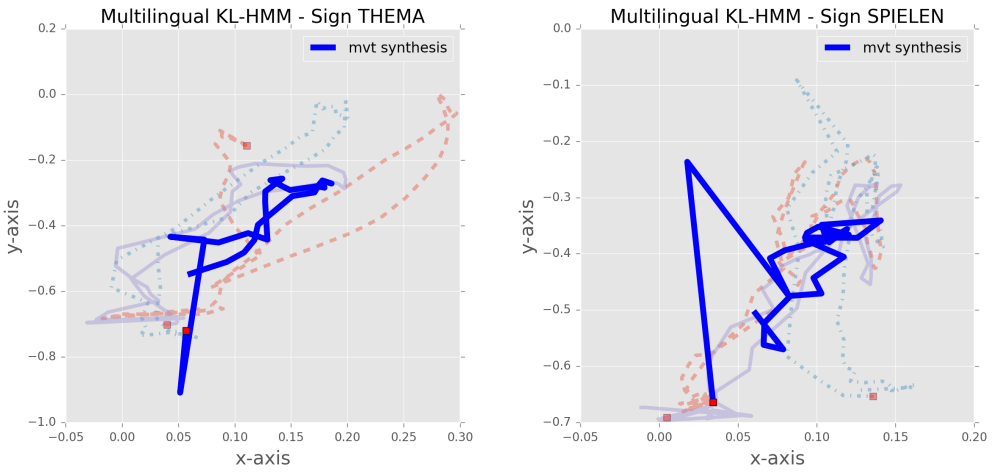

Figure 5.6 – Hand movement synthesis of the dominant hand for the well-recognized sign THEMA (left) and the poorly-recognized sign SPIELEN (right) using the Gaussian distribution sequence computed from the KL-HMM system using the TSL subunits. The red squares are the starting points.

5.4.3 Impact of number of training samples

As discussed in Chapter 1, sign languages are inherently under-resourced. To handle that, the cross-lingual approach is leading towards methods to share resources across sign languages. So, we investigated the impact of number of training samples per sign from the target sign language on the performance of cross-lingual sign language recognition system.

In the SMILE DSGS database, since we only used Category 1 an Category 2 data (see Section 2.3), the number of training samples varies; Figure 5.7 depicts the histogram of the number of samples per sign. To find the appropriate number of training samples per sign, we conducted a study using three different setups where in the first setup ten samples per sign (referred as *ten-sample-signs*), in the second setup eight samples per sign (*eight-sample-signs*) and in the third setup six samples per sign (*six-sample-signs*) are used. To evaluate the three different setups, we derived hand movement subunits for each setup and built KL-HMM based sign language recognition systems.

Table 5.3 presents the RA obtained on the three setups. It can be observed that the RA decreases with decrease in number of samples per sign. As it can be observed in Figure 5.8, the minimum number of samples seems to be around twelve samples but, the figure also shows that the number

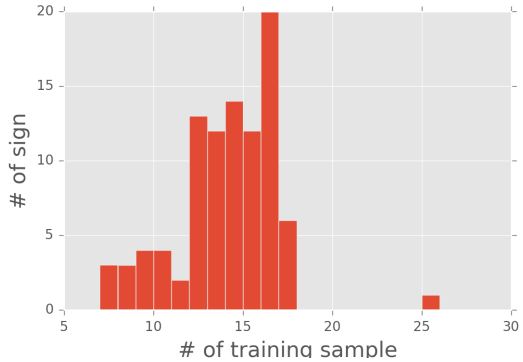

Figure 5.7 – Histogram of the number of training samples per sign of the SMILE DSGS database.

Table 5.3 – Cross-lingual KL-HMM based systems results on the SMILE DSGS database depending on the three different setups used to infer the lexicon (*ten-/eight-/six-sample-signs lexicon*)

Number of training samples	Sign RA
ten-sample-signs	35.9
eight-sample-signs	33.7
six-sample-signs	33.8

of samples needed depends on the sign. The different movement complexity of the signs and variations introduced by the signers can explain this difference. Furthermore, as we observed earlier, the differences in the coverage of hand movements across the two databases can also explain why a low number of samples is not sufficient.

5.5 Model Selection-based Sign-level HMM Inference for Subunits Extraction

In Step 2 of the proposed approach (see Section 5.2) a sign level HMM is inferred for each sign. The HMM structure is defined as left-to-right HMM with a fixed number of states for all the signs based on the recognition accuracy obtained on the training and the development data. In Chapter 4, we developed a model selection approach which can be applied to define the HMM structure in a data driven manner. In this section, we apply that approach in the proposed subunit extraction framework, by implementing it in Step 2.

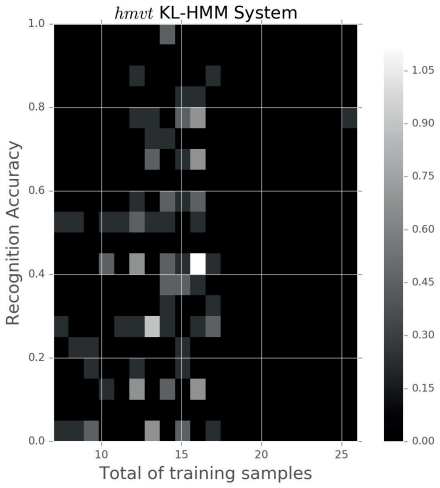

Figure 5.8 – Sign recognition accuracy density of the KL-HMM system per sign using the TSL hand movement subunit according to the total number of training samples.

5.5.1 Proposed modification

In the proposed modification, in Step 2, a range of possible number of states: $[N_{min}, N_{max}]$ is first defined. Then, an HMM with n states is modeled for each sign $S^m \in \{S^1 \cdots S^M\}$, $\forall n \in [N_{min}, N_{max}]$. Finally, from the set of HMM for each sign, the HMM that yields the maximum likelihood on the development data is chosen. Figure 5.9 illustrates this process. Given the

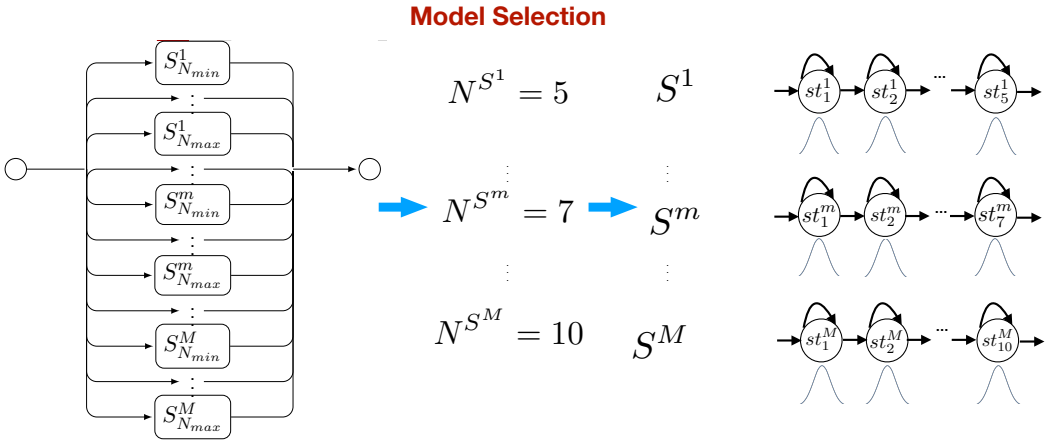

Figure 5.9 – Illustration of the *MS-based* sign level HMM topology inference

inferred HMM with N^{S^m} states for each sign S^m, the resulting sign model is a sequence of Gaussian distributions. Step 3 (see Section 5.2) remains the same, i.e., the HMM states are clustered through the same measure of discrimination, i.e. Bhattacharya distance.

5.5.2 Experimental setup

We compared the proposed modification in Step 2 with the former method of inferring sign level HMM based on the recognition accuracy on the training and development data on signer-independent SLR and gesture recognition tasks. For the sake of clarity, we denote the former method as *std-based* and the method with proposed modification as *MS-based*.

Databases

The HospiSign sign language database and the Chalearn14 gesture database were used in this study; more details can be found in Section 2.3.

In the HospiSign database, in order to conduct a signer-independent experiment, we have used a leave-one-signer out cross-validation study. Furthermore, as we need a development set for tuning the hyper-parameters, we have left another signer out for this purpose. Therefore, as can be seen from Table 5.4, we have conducted six experiments where in each experiment, one signer is used for testing, one signer is used as the development set, and the rest of the signers are used for training. For each experiment, we have presented the average performance over the signers as the final result. Table 5.5 presents the average number of samples in the train, development and test sets over the six experiments.

Table 5.4 – The HospiSign database segmentation of training, development and testing data according to signers. The numbers in the table refer to the signers's number

	Exp 1					Exp 2					Exp 3					Exp 4					Exp 5					Exp 6				
Train	3,4, 5,6	2,4, 5,6	2,3, 5,6	2,3, 4,6	2,3, 4,5	3,4, 5,6	1,4, 5,6	1,3, 5,6	1,3, 4,6	1,3, 4,5	2,4, 5,6	1,4, 5,6	1,2, 5,6	1,2, 4,6	1,2, 4,5	2,3, 5,6	1,3, 5,6	1,2, 5,6	1,2 ,3,6	1,2, 3,5	2,3, 4,6	1,3, 4,6	1,2, 4,6	1,2, 3,6	1,2, 3,4	2,3, 4,5	1,3, 4,5	1,2, 4,5	1,2, 3,5	1,2, 3,4
Dev	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3	4	4	4	4	4	5	5	5	5	5	6	6	6	6	6
Test	2	3	4	5	6	1	3	4	5	6	1	2	4	5	6	1	2	3	5	6	1	2	3	4	6	1	2	3	4	5

Table 5.5 – Description of the HospiSign database in terms of average number of samples

	Train	Dev	Test
# of samples	874	210	210

For efficient comparison to the existing results, we used the train/development/test setup defined by the Chalearn14 competition. Table 5.6 presents the number of samples in the train, development and test sets. It is worth mentioning that in the Chalearn 2014 competition, the segmentation of

Table 5.6 – Description of the Chalearn14 database

	Train	Dev	Test
# of samples	6800	2506	3579

videos in the test set was not provided. Therefore, the task contained two parts: (1) segmentation of videos into gestures/non-gestures, and (2) classifying the gestures. As our focus in this section is on classifying the gestures, we have used the ground truth segmentation on the test set.

We are aware that signs and gestures are different but, for sake of simplicity, we used the term *sign* to refer to sign and gesture in the remainder of this section.

Systems

We built a HMM/GMM and a hybrid HMM/ANN systems using sign level sequence modeling approach and the SU-based sequence modeling approach.

HMM/GMM Systems:
In the case of sign level sequence modeling, we modeled the signs with left-to-right HMM using Gaussian state-output distributions. The number of Gaussian mixtures varies between 1 and 56 and was defined by the best recognition accuracy on the development data. We defined both the *std-based* and the *MS-based* inferences to determine the number of HMM states per sign. The range of HMM states was from $N_{min} = 3$ to $N_{max} = 9$ states. In the *MS-based* framework, the average number of HMM states per sign was 8 for the HospiSign database and 3 for the Chalearn14 database. In the *std-based* framework, the derived number of HMM states per sign was defined such that the recognition accuracy saturates on the training data; the resulting number was 9 for both databases.

In the SU-based model, we trained HMM/GMM systems where each subunit was modeled with a single HMM state. The number of Gaussian components per mixture also varied between 1 to 56, and was set based on the recognition accuracy on the development set.

Hybrid HMM/ANN Systems:
For building the hybrid HMM/ANN systems, we first obtained the alignments in terms of the HMM states using the trained HMM/GMM systems. We then trained ANNs, more precisely multilayer perceptrons (MLPs) classifying HMM states with output non-linearity of softmax and minimum cross-entropy error criterion. We used 36-dimensional position and velocity features with four frames preceding context and four frames following context as the MLP input. In our experiments we trained MLPs with different number of hidden units (600, 800, 1000) and hidden layers (0, 1, 2, 3). The number of hidden units and hidden layers as well as other hyper-parameters such as learning rate and the batch size were chosen according to the frame-level accuracy on the development set.

We estimated the scaled likelihoods in the hybrid HMM/ANN systems by dividing the posterior probabilities derived from MLPs with the prior probabilities of the classes estimated from relative

frequencies in the training data. These scaled likelihoods were then used as emission probabilities for HMM states.

The performance of the developed systems are evaluated in terms of recognition accuracy (RA) described in Section 2.4.

5.5.3 Results and analysis

In this section, we first present the recognition accuracies on the HospiSign and Chalearn14 databases. We then contrast the performance of the proposed approach with the existing approaches in the literature

Sign language recognition on HospiSign database

Table 5.7 presents the HMM/GMM SLR system results (± standard deviation) in terms of *RA* on the HospiSign database; the detailed signer split for the presented experiment setup can be found in Table 5.4.

Table 5.7 – HMM/GMM results on the HospiSign database depending on the *std-based* and the *MS-based* segmentation approach explained in Section 5.5.1

Experiment	sign level seq. modeling		SU-based seq. modeling	
	std-based	*MS-based*	*std-based*	*MS-based*
Exp 1	92.6	92.0	92.0	90.8
Exp 2	89.5	89.7	87.9	88.0
Exp 3	91.7	92.9	92.1	90.9
Exp 4	88.1	89.5	89.5	89.7
Exp 5	91.6	90.0	91.6	91.2
Exp 6	93.0	92.0	90.6	88.3
Average *RA* ± std	91.1 ± 1.9	91.0 ± 1.4	90.6 ± 1.7	89.8 ± 1.4
Average # of states	*300*	*278*	*218*	*190*

Firstly, it can be observed that the SU-based modeling approach leads to development of a comparable SLR system to the sign level sequence modeling approach; thus we confirm what was observed with SMILE DSGS database in Section 5.3. This observation also holds for *MS-based* systems. This is interesting as the *MS-based* SU-based system is able to perform comparable to the *MS-based* sign level system despite considerably reducing the total number of states by in average 32%. Secondly, the *MS-based* system yields comparable recognition accuracy in the sign level setup while in the SU-based setup there is slight decrease in performance. Altogether, it can be observed that if both model selection approach and subunit clustering is applied we can reduce

the number of states from 300 to 190, i.e. around 37%, while keeping comparatively similar good recognition accuracy.

Table 5.8 presents the hybrid HMM/ANN results on the HospiSign database. It can be observed that the use of neural networks instead of GMMs for estimating the local emission scores leads to significant improvements in the performance of all the systems.

Table 5.8 – Hybrid HMM/ANN results on HospiSign database depending on the *std-based* and the *MS-based* segmentation approach explained in Section 5.5.1

Experiment	sign level seq. modeling		SU-based seq. modeling	
	std-based	*MS-based*	*std-based*	*MS-based*
Exp 1	96.6	94.8	96.9	94.5
Exp 2	96.0	95.1	95.1	94.1
Exp 3	96.5	97.0	95.9	96.2
Exp 4	95.5	95.8	95.1	94.5
Exp 5	94.3	94.0	95.0	96.5
Exp 6	95.1	94.8	95.4	94.5
Average *RA* ± std	95.7 ± 0.9	95.2 ± 1.0	95.6 ± 0.7	95.0 ± 1.0

Comparison of the results on the six experimental setups depicted in Table 5.4 shows that, irrespective of the development set chosen, the systems perform mostly similar to one another. This indicates that the proposed subunit extraction approach is less sensitive to these aspects.

Gesture recognition on Chalearn14 database

Table 5.9 presents the HMM/GMM and hybrid HMM/ANN results on Chalearn14 database. In the HMM/GMM systems based on the gesture level sequence modeling, the average number of Gaussian mixtures used is 40. Indeed, the number of Gaussian mixtures plays an important role in the performance of the systems as increasing the number of mixtures from 1 to 40 leads to around 25% absolute improvement in the gestures recognition accuracy. This improvement can be explained by the wild setup and the gesture framework which imply a significant signer variation. So, the balance between the number of states and the number of mixtures is more difficult to set compared to HospiSign study where increasing the number of mixtures does not change the recognition accuracy. In the Chalearn14 case, the SU-based HMM/GMM model seems to better handle this balance since we can notice a significant improvement compared to the gesture level model. Furthermore, in the *std-based* setup, the SU-based system improves over the gesture level sequence modeling with a decrease of around 14% of number of states, while in the *MS-based* setup it is not the case. This can be explained by the fact that the *MS-based* setup needs a splitting criteria since the segments are not minimal, whereas, such a criteria is not

needed for the *std-based* setup.

Table 5.9 – HMM/GMM and hybrid HMM/ANN results on Chalearn14 database depending on the *std-based* and the *MS-based* segmentation approach explained in Section 5.5.1

System	gesture level seq. modeling		SU-based seq. modeling	
	std-based	*MS-based*	*std-based*	*MS-based*
HMM/GMM	80.8	83.5	86.1	78.6
Hybrid HMM/ANN	81.3	83.0	83.8	78.5
Average # of states	*183*	*70*	*157*	*55*

In conclusion, the results suggest that it is somewhat better to infer the subunit-based on the *std-based* segmentation rather than the *MS-based* segmentation. Furthermore, *std-based* SU extraction tends to yield recognition systems comparable to sign- or gesture-level sequence modeling-based systems.

Comparison to existing studies

In this section, in order to ascertain that our approach is leading to meaningful systems, we contrast our results with the performance of systems reported on HospiSign and Chalearn14 databases which use the skeleton information.

Comparison on HospiSign database

In [18], various manual features such as handshape, hand position and hand movement were extracted and temporal modeling using either DTW or temporal templates was performed. In the case of using DTW, the signs were classified using k-Nearest Neighbors. In the case of using temporal templates, Random Decision Forest was used for classifying the signs.

For a fair comparison, as done in [18], we first computed the average performance for each signer over the six experiments, and then computed the average performance over the signers as the final accuracy. Table 5.10 provides the comparison of our approach using the hybrid HMM/ANN framework with the proposed approach in [18] when using DTW along with k-Nearest Neighbors using hand joint distances and hand movement distances as features. Furthermore, we have presented the results in the case of using temporal templates with the random decision forests as it yielded one of the best results in [18]. It can be observed from Table 5.10 that both sign-level and SU-based sequence modeling approaches yield comparable systems to the systems developed in [18]. Furthermore, the lower standard deviation w.r.t DTW & k-Nearest Neighbors based systems indicates that the proposed approach is yielding a more consistent system across different

signers.

Table 5.10 – Comparison of our proposed approach to the proposed approach in [18] for the HospiSign database

	Approach	Features	Accuracy
Our approach	HMM/ANN with *std-based* sign level seq. modeling	position and motion	95.7 ± 2.5
Our approach	HMM/ANN with *std-based* SU-based seq. modeling	position and motion	95.6 ± 2.8
Approach in [18]	DTW & k-Nearest Neighbors using hand movement distance	position and motion	93.8 ± 6.4
Approach in [18]	Temporal templates & Random Decision Forest	position, motion and handshape	96.7 ± 1.8

Comparison on Chalearn14 Database

In the Chalearn 2014 competition, various approaches for feature extraction, temporal segmentation and classification of gestures were investigated. In order to evaluate the proposed approaches, Jaccard index was used as the evaluation metric. Jaccard index is a commonly used metric for evaluating the gesture spotting. The Jaccard index is defined as:

$$J_{s,g} = \frac{A_{s,g} \cap B_{s,g}}{A_{s,g} \cup B_{s,g}}, \tag{5.2}$$

where $A_{s,g}$ is the ground truth for gesture g at sequence s, and $B_{s,g}$ is the prediction for this gesture at sequence s [38].

Table 5.11 contrasts the performance of the systems based on the proposed subunit extraction approach with the performance of the systems in the competition that used only the skeleton information, like the proposed approach. It is worth mentioning that in the Chalearn 2014 competition, the segmentation of videos in the test set was not provided. Therefore, the task contained two parts: (1) segmentation of videos into gestures/non-gestures, and (2) classifying the gestures. As our focus is on classifying the gestures, we have used the ground truth segmentation on the test set. In order to get an idea on how the systems resulting from the proposed approach perform when the ground truth information is not available, we evaluated our systems based on the segmentation used in the system reported in [17].[2] When considering segmentation and classification, we can observe that the systems based on the proposed approach are neither the

[2] In [17] a random forest was used to recognize the gestures and non-gestures. We would like to thank Necati Cihan Camgöz for sharing the test set segmentation with us.

best nor the worst. Thus, indicating that the proposed approach is worth pursuing for gesture recognition as well.

Table 5.11 – Comparison between the performance of the proposed approach with the performance of related approaches from the Chalearn 2014 competition in terms of Jaccard index

Team/Approach	Accuracy	Features	Classifier
std-based SU-based seq. modeling (ground truth seg.)	0.8655	Skeleton	HMM/GMM
MS-based gesture level seq. modeling (ground truth seg.)	0.8422	Skeleton	HMM/GMM
Ismar [17]	0.7466	Skeleton	Random forest
std-based SU-based seq. modeling (seg. from [17])	0.6868	Skeleton	HMM/GMM
MS-based gesture level seq. modeling (seg. from [17])	0.6825	Skeleton	HMM/ANN
Terrier	0.5390	Skeleton	Random forest
YNL	0.2706	Skeleton	HMM, SVM

5.6 Discussion and Summary

The present chapter proposed a data-driven approach for hand movement subunit extraction for modeling signs and gestures without using any linguistic annotation information. Specifically, the subunits are extracted given only pairwise comparison between each pair of sign productions or gesture productions. As it can be seen in Table 5.12, the previous approaches have focused on processing images or motion information captured via gloves, while our approach focuses on modeling skeleton information, which can be easily and reliably obtained nowadays. Also, most of these works have not investigated signer-independence. Furthermore, we also demonstrated that the extracted subunits could be transferred across different sign languages. An aspect that has not been studied by the existing approaches in the literature.

Besides that, the recognition studies consistently showed that, with the 36 dimensional hand movement features extracted from the skeleton information, the subunits can be obtained at the sign-level itself. Clustering of the sign-level HMM states through pairwise discrimination leads to state space reduction. In both scenarios, discrimination between the signs gets modeled and leads to systems that yield similar performance. The investigations also showed that the proposed approach is not so sensitive to the method used to infer sign-level HMM in Step 2, i.e. *std-based* or *MS-based*. Both methods lead to extraction of subunits that yield comparable recognition performance. Finally, through development of a visualization method inspired from robotics, we

Table 5.12 – Comparison of our sign language hand movement subunit studies with existing studies

Ref.	Features based	Segment.	Clustering algorithm	Recognition study	Signer indep. study	Monolingual/ Cross-lingual
Sako and Kitamura [98]	images processing	multi-stream HMM	tree based algorithm	✓	✓	Monolingual
Bauer and Kraiss [8]	gloves	HMM	k-means	✓	✗	Monolingual
Han et al. [53]	images processing	discontinuity detector	DTW	✓	✗	Monolingual
Fang et al. [42]	gloves	HMM	modified k-means	✗	✗	Monolingual
Theodorakis et al. [111]	images processing	HMM	HMM hierarchical clustering	✗	✗	Monolingual
Our approach	skeleton	HMM	pair-wise clustering with Bhatt. dist.	✓	✓	Mono- and cross-lingual

demonstrated that the extracted subunits could be further analyzed through synthesis of the hand movements of signs in the input 3D coordinate space.

In the following chapters, we build upon the proposed subunit extraction approach to develop methods to model hand movement information with handshape information for sign language recognition and sign language assessment.

6 Phonology-based Sign Language Recognition Framework

RQ: How to model the multichannel information inherent to sign languages?

A significant progress was achieved for sign languages in the 60s with the linguistic analysis of American Sign Language (ASL) by William Stokoe [105]: sign language has his own phonology, morphology and syntax. The phonology describes the organisation of non-meaningful distinctive units of language (phonemes) which combine themselves to form meaningful units (morphemes). In spoken languages, the phonemes are acoustic and articulatory-based units. In sign languages, the phonology is visual-based: the phonemes are manual and non-manual components, such as handshapes, hand trajectories, hand localisation and mouthing, which combines themselves to produce new vocabulary signs. A phoneme substitution to another allows to distinguish two signs, as for example in Swiss French Sign Language (Langue des Signes Française) (LSF) the signs MARS (march) and DIRE (to say) differ only by the hand movement (see Figure 6.1). One main difference with spoken languages is that the unit combination is simultaneous while in spoken language it is sequential. Thereby modeling the multichannel information of a sign is a highly challenging problem. In this chapter, we propose a phonology-based framework to model sign languages by elucidating the link between spoken language and sign language in terms of production phenomenon and perception phenomenon. Specifically, we elucidate that when modeling linguistically motivated speech production knowledge, i.e. "articulatory" features (AFs), it is a multichannel information modeling problem akin to sign language processing. Through that understanding, we show that the methods developed to model "Articulatory" Features (AF)s can be scaled to model the multichannel information for sign language processing.

Moreover, modeling of multichannel information requires also sufficient sign language specific data. This is a challenge as sign languages are inherently under-resourced, as discussed in

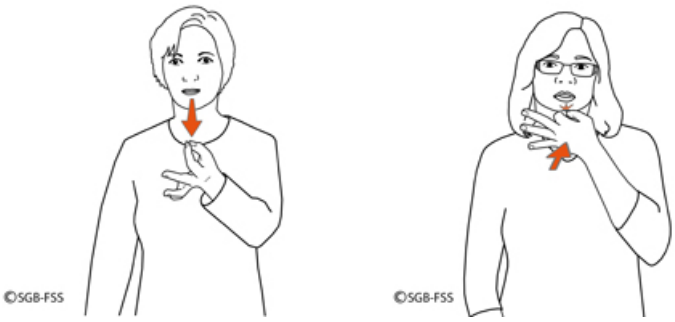

Figure 6.1 – Illustration of the sign production of the Swiss LSF signs DIRE (left) and MARS (right) ("to say" and "March", respectively).

Chapter 1. One way to address the resource scarcity challenge is to develop methods that can exploit multiple sign language resources by overcoming the limitations imposed by the differences between the sign languages. In the literature, there is limited work in that direction, more precisely with handshape modeling only. It has been found that, given the HamNoSys annotation [45] of produced signs, a global handshape classifier can be trained by pooling resources from multiple sign languages and handshape information based sign language recognition systems can be developed [60]. However, handshape is only one channel of information. There is need to model other channels such as, hand movement, which unlike handshape is a continuous aspect or in other words is not inherently a discrete unit. In Chapter 5, we demonstrated that the hand movement information can be discretized into subunits and these subunits exhibit language-independent characteristics. We build upon that to demonstrate that, in the proposed phonology-based frameworks to model multichannel information, sign language recognition systems can be effectively developed by using multilingual sign language resources.

The remainder of the chapter is organized as follows: in Section 6.1, we present the related work. Section 6.2 presents the proposed phonology-based approaches to jointly the model multichannel information inherent in sign languages. Section 6.3 and Section 6.4 validate the proposed phonology-based approaches through monolingual and multilingual studies, respectively. Finally, Section 6.5 concludes the chapter with a summary of key findings. The material presented in this chapter is largely based on the following publications:

> *HMM-based approaches to model multichannel information in sign language inspired from articulatory features-based speech processing*, Sandrine Tornay, Marzieh Razavi, Necati Cihan Camgoz, Richard Bowden and Mathew Magimai.-Doss, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019

Towards multilingual sign language recognition, Sandrine Tornay, Marzieh Razavi and Mathew Magimai.-Doss, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020

6.1 Related Work

Sign language processing presents two main challenges: (1) robust extraction of the multichannel information and (2) modeling of the multichannel information. In this chapter, we are addressing the second challenge, i.e. modeling of the multichannel information. The first one was addressed in Chapter 3 and Chapter 5 where initially the handshape estimators and the hand movement extraction were presented (Chapter 3) and then the method to extract the hand movement subunits (Chapter 5).

As presented in Section 2.2, different machine learning techniques have been investigated for modeling signs for SLR such as, HMM [104], parallel HMM [117], relevance vector machines [122] and deep learning methods [62, 20]. The early work of Vogler and Metaxas [115] borrowed heavily from the studies of sign language by Liddell and Johnson [71], splitting signs into motion and pause sections. While their later work [117], used parallel HMM on both handshape and hand motion subunits, as proposed by the linguist Stokoe [105].

Sign language processing faces data scarcity issues. Thus, the studies have also concentrated on learning sign models in an effective manner from low number of examples. Lichtenauer et al. [69] presented a method to automatically construct a sign language classifier for a previously unseen sign. Their method works by collating features for signs from many people then by comparing the features of the new sign to that set. They then construct a new classification model for the target sign. This relies on a large training set for the base features (120 signs by 75 people) yet subsequently allows a new sign classifier to be trained using one shot learning. Bowden et al. [15] also presented a SLR system capable of correctly classifying new signs given a single training example. Their approach used a two-stage classifier bank, the first of which used hard coded classifiers to detect handshape, hand arrangement, motion and position "subunits". The second stage removed noise from the 34 bit feature vector (from stage 1) using independent component analysis, before applying temporal dynamics to classify the sign. Kadir et al. [54] extended this work with head and hand detection based on boosting (cascaded weak classifiers), a body-centered description (normalized movements into a 2D space) and then a two-stage classifier where stage 1 classifier generates linguistic feature vector and stage 2 classifier uses Viterbi on a Markov chain for highest recognition probability. Cooper and Bowden [27] continued this work still further with an approach to SLR that does not require tracking. Instead, a bank of classifiers is used to detect "phonemic" parts of sign activity by training and classifying (AdaBoost cascade) on certain

sign subunits. These were then combined into a second stage word-level classifier by applying a first order Markov assumption. The results showed that the detection rates achieved with a large lexicon and few training examples were almost equivalent to a tracking-based approach. With the advances in deep learning methods, there has been effort in modeling signs in the framework of hybrid HMM/ANN [62] and in the framework of connectionist temporal classification [20]. However, these efforts have mainly focused on modeling handshape information.

6.2 Proposed Phonology-based Framework for Sign Language Recognition

Conceptually, communication is about transmission of a signal between a source and a receiver, see Figure 6.2. The source produces a signal and the receiver perceives it. For example, in speech communication a human source produces speech signal by moving the articulators in the speech production system and a human receiver perceives the signal as sequence of phones, words and sentences/phrases. In sign language communication, a human produces visual signal through manuals and non-manuals and a human receiver perceives them as words and phrases. In other words, communication involves a synergy between production phenomenon and perception phenomenon.

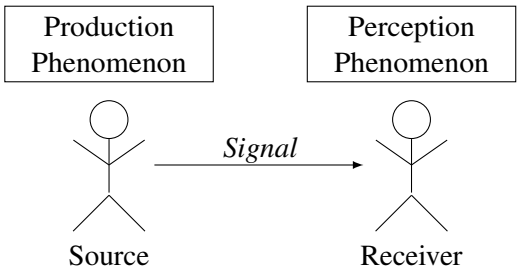

Figure 6.2 – Illustration of communication scheme.

More precisely in both sign language and spoken language:

(a) there is a production phenomenon that generates a signal. In the case of spoken language, it is movement of articulators like vibration of vocal folds, movement of tongue, lips and jaw that produce time varying 1D acoustic signal. In the case of sign language, it is hand articulators (such as handshape), mouthing, body postures and facial expressions that produce time varying 2D visual signal; and

(b) there is a perception phenomenon, which interprets that generated signals in terms of elements of “language”, e.g. words, phrases.

Linguistically, the perception phenomenon is better understood in spoken language than sign language. More precisely, in spoken language, it is well understood that the time structure of word units can be defined as a sequence of subword units, e.g. phonemes, syllables, which are "perceptual" in nature (i.e. can be heard and distinguished); can be related to the movement of articulators; and can be modelled by parameterizing the spectral characteristics of the speech signal. Such an understanding, however, does not exist yet in the case of sign language. More precisely, how hand gestures, facial expressions, body postures, mouthing together create a subword unit like a time structure is not clear yet. It is still an open research problem in sign linguistics.

In spoken language processing, despite the success of spectral feature-based approach, there is interest in modeling the production phenomenon related information through AFs [58, 73, 94]. More precisely, defining each phoneme in terms of AFs like manner of articulation or degree of constriction, place of articulation, voicing, nasality, rounding, height of tongue, frontness of tongue; estimating these AFs from the speech signal; and then modeling the multichannel AFs through sequential models such as HMM. The AFs in speech processing are synonymous to the "subunits" in sign language. This close similarity can be exploited to scale the methods developed for AF based processing to sign language processing. In this chapter, we develop two such methods, namely, an HMM/GMM-based approach which models tandem features and a KL-HMM-based approach that models posterior probabilities of visual subunits.

6.2.1 HMM/GMM-based approach

One of the common approach to model AFs is to estimate these features using ANN; transform them using tandem feature extraction technique; concatenate them with the acoustic feature; and model them with HMM [73, 40, 22, 94]. As illustrated in Figure 6.3, we can adopt a similar approach for sign language processing where the features representing different channels of information are extracted, concatenated $\mathbf{x}_t := \left[\mathbf{x}_t^{\text{hshp}}\ \mathbf{x}_t^{\text{hmvt}}\ \cdots\ \mathbf{x}_t^{\text{facial}}\right]^T$ and then modeled by an HMM/GMM. $\mathbf{x}_t^{\text{hshp}}$, $\mathbf{x}_t^{\text{hmvt}}$, $\mathbf{x}_t^{\text{facial}}$ denote the features corresponding to handshape, hand movement and facial expression, respectively. The features can be extracted in the measurement space as for example the hand movement features $\mathbf{x}_t^{\text{hmvt}}$ by using the 3D skeleton features or from the probabilistic representation of the subunits using tandem technique [47], as for example the handshape features. The tandem feature extraction technique is described in Section 3.3.1.

6.2.2 Kullback-Leibler divergence HMM-based approach

Another approach is to model AFs as probabilistic features using KL-HMM (see Section 3.3.2). As illustrated in Figure 6.4, we can adopt the KL-HMM based AF modeling framework that was

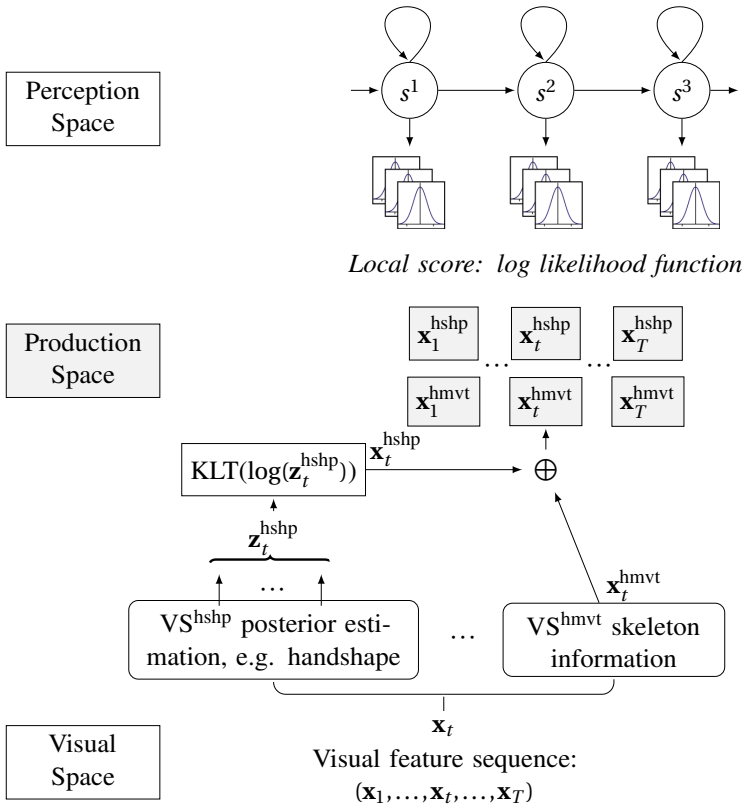

Figure 6.3 – Illustration of the tandem feature-based HMM/GMM approach to model multichannel information for sign language processing; VS stands for Visual Subunits.

originally proposed in [94] for sign language processing, where for each channel we extract probabilistic features and stack them to get the feature observation $\mathbf{z}_t := \left[\mathbf{z}_t^{\text{hshp}}\ \mathbf{z}_t^{\text{hmvt}}\ \cdots\ \mathbf{z}_t^{\text{facial}}\right]^T$. $\mathbf{z}_t^{\text{hshp}}$, $\mathbf{z}_t^{\text{hmvt}}$ and $\mathbf{z}_t^{\text{facial}}$ denote the probabilistic features corresponding to handshape, hand movement and facial expression, respectively. The HMM state s^i is parameterized by a stack of categorical distribution $\mathbf{y}_{s^i} := \left[\mathbf{y}_{s^i}^{\text{hshp}}\ \mathbf{y}_{s^i}^{\text{hmvt}}\ \cdots\ \mathbf{y}_{s^i}^{\text{facial}}\right]^T$ of the same dimension as the feature observations. The local score $S(\mathbf{y}_{s^i}, \mathbf{z}_t)$ is based on KL-divergence [65].

6.3 Monolingual Sign Language Recognition

We validated both the proposed approaches on a monolingual isolated sign language recognition (SLR) task using the SMILE DSGS database.

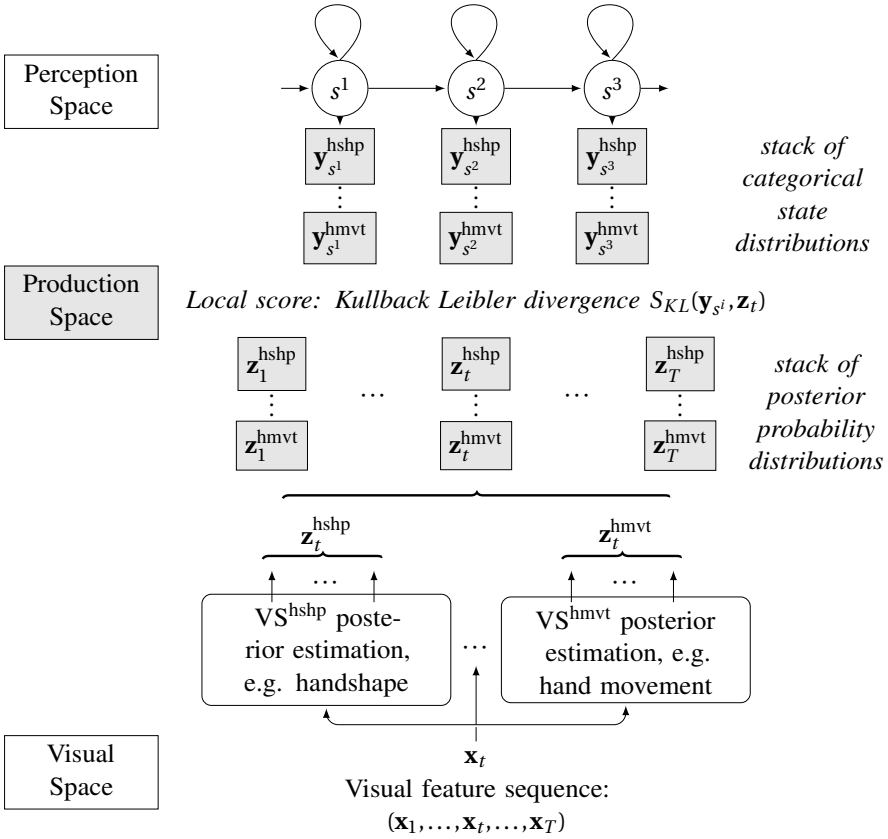

Figure 6.4 – Illustration of the KL-HMM approach to model multichannel information for sign language processing; VS stands for Visual Subunits.

6.3.1 Database

As mentioned, the large-scale SMILE Swiss German sign language database presented in Section 2.3 was used. We only used, in our experimental setup, the second pass annotated as Category 1 or 2, as there are acceptable signs with the same or slightly the same form. The data was partitioned in a signer-independent manner into 1263 training set samples from 17 signers, 249 development set samples from 3 signers and 704 test set samples from 10 signers.

6.3.2 Feature estimation

Handshape features

In the KL-HMM approach, as presented in Section 3.1.2, the off-the-shelf DeepHand neural network was used to estimate the handshape class-conditional posterior probabilities, $\mathbf{z}_t^{\text{hshp}}$.

In the HMM/GMM approach, the $\mathbf{z}_t^{\text{hshp}}$ was transformed into tandem feature $\mathbf{x}_t^{\text{hshp}}$ by applying the logarithm and Kahunen Loeve Transform (KLT) operations. In the KLT step, we reduced the feature dimension to cover up to 99% variance of the data.

Hand movement features

The hand movement feature observations $\mathbf{x}_t^{\text{hmvt}}$, used in the HMM/GMM approach, are 36 dimensional the position and velocity skeleton-based features presented in Section 3.1.1.

In the KL-HMM approach, we used the hand movement subunit extraction approach presented in Chapter 5, where the sign level HMM was used to model $\mathbf{x}_t^{\text{hmvt}}$ into GMMs. We then estimated the posterior probability-based features $\mathbf{z}_t^{\text{hmvt}} := [z_t^1, \ldots, z_t^I]^T$ with the following two methods:

(1) by applying the Bayes' rule on the GMMs;

(2) with MLPs, by aligning $\mathbf{x}_t^{\text{hmvt}}$ in terms of the HMM states and then trained MLPs classifying HMM states with output non-linearity of softmax and minimum cross-entropy error criterion. We used the feature observation with four frames preceding context and four frames following context as the MLP input. In our experiments, we trained MLPs with different number of hidden units (600, 800, 1000) and hidden layers (0, 1, 2, 3). The number of hidden units and hidden layers as well as other hyper-parameters such as learning rate and the batch size were chosen according to the frame-level accuracy on the development set. The MLPs were trained using the Quicknet software [52].

The total number of HMM states is $I = 849$.

6.3.3 Recognition models

We built three systems for each of the two proposed approaches: the hand movement-based system (**M**), the handshape-based system (**rlS**) and the handshape-plus-hand movement-based system (**M+rlS**).

Following Chapter 4, in both approaches, the number of states in the left-to-right HMM where varied from 3 (N_{min}) to 9 (N_{max}) during training. So each sign had 7 different HMM. During recognition phase, the decoder selected from 94×7 sign models the sign model that yielded the maximum likelihood in the case of HMM/GMM approach and the sign model that yielded the minimum KL-divergence score in the case of KL-HMM approach. For the HMM/GMM approach, each state was modeled by 4 Gaussian mixtures for the **M** system and by a single Gaussian for systems **S** and **M+S**. We found that increasing the number of mixture of Gaussians for **S** and **M+S** did not help in improving performance.

6.3.4 Results

Table 6.1 shows the performance obtained in terms of recognition accuracy for both proposed approaches: the HMM/GMM approach and the KL-HMM approach. We report the performance according to the number of states, i.e. developing the system by presetting the number of HMM states. **ms 3 to 9** denotes the model selection approach based system, where the HMM topology is inferred during decoding (as presented in Chapter 4).

Table 6.1 – Recognition accuracy of the HMM/GMM and the KL-HMM approaches applied on the hand movement (hmvt) features (**M**), the handshape features (**S**) and combined ones (**M+S**).

	HMM/GMM			**KL-HMM**				
				hmvt estimation based on GMM			hmvt estimation based on MLP	
#state	**S**	**M**	**M+S**	**S**	**M**	**M+S**	**M**	**M+S**
3	47.7	44.4	63.8	25.9	41.5	59.5	43.8	68.3
4	47.6	47.2	63.4	28.8	39.9	60.5	46.7	70.5
5	49.3	48.5	64.8	28.0	41.8	60.1	45.6	69.7
6	45.3	49.8	65.5	28.0	43.0	62.4	46.3	71.0
7	46.6	48.1	66.1	30.7	41.2	60.5	49.6	69.9
8	44.6	50.2	63.7	32.5	43.3	62.1	47.0	70.0
9	43.5	50.4	65.9	30.7	41 9	61.7	48.0	71.7
ms 3 to 9	50.3	51.6	**66.8**	32.8	44.3	**63.1**	47.3	**71.9**

It can be observed that, in both the approaches, **M+S** systems outperform handshape alone and hand movement alone systems. The model selection method **ms 3 to 9** yields the best system for both the approaches. When comparing across the approaches, the HMM/GMM approach yields better system than KL-HMM. Low performance for system **S** in KL-HMM approach can be explained from the fact that the DeepHand handshape posterior feature estimator has not observed any SMILE DSGS dataset. However, the HMM/GMM approach uses SMILE training data to get the KLT matrix. For system **M** in KL-HMM approach the MLP based posterior estimation gives better performance compare to the GMM based posterior estimation. Since the HMM/GMM **M** system yields the best performance, we suppose that standard HMM/GMM is sufficient to model the hand movement information alone while in the combined handshape and hand movement features (system **M+S**), KL-HMM approach with MLP based posterior estimation outperforms the HMM/GMM approach.

6.3.5 Analysis

This section presents the advantage of the interpretability of the KL-HMM approach through two analyses: one on the hand movement channel and one on the handshape channel.

Hand position and velocity decomposition analysis

The proposed approaches, in particular KL-HMM approach, allows further simplifications. For instance, the hand movement can be decomposed into position and velocity and can be modeled independently with handshape. We demonstrate that through an experiment with KL-HMM approach. We used the GMM-based posterior estimation method used to derive the hand movement posterior feature estimator (see section 6.3.2) to obtain the hand position $\mathbf{z}_t^{\text{hpos}}$ and the hand velocity $\mathbf{z}_t^{\text{hvel}}$ posterior feature estimators based on the hand position and velocity features separately. Table 6.2 presents the results. System **P** denotes modeling of $\mathbf{z}_t^{\text{hpos}}$ alone. System **V** denotes modeling of $\mathbf{z}_t^{\text{hvel}}$ alone. **P+S** and **V+S** denotes modeling of handshape posterior feature $\mathbf{z}_t^{\text{hshp}}$ along with $\mathbf{z}_t^{\text{hpos}}$ and $\mathbf{z}_t^{\text{hvel}}$, respectively. We can observe the same trends as before that jointly modeling handshape and hand position or hand velocity information helps. It can be observed that separating the hand movement features into position and velocity (the **P+V** system) does not affect the performance in comparison to the **M** system. Furthermore, we see that there is a slight increase in the performance of system **V+S** when compared to system **M+S**.

Table 6.2 – Recognition accuracy of the KL-HMM approach applied on the hand position features (**P**), the hand velocity features (**V**), both features (**P+V**), and each combined with the handshape features (·**+S**)

	KL-HMM					
#state	**P**	**V**	**P+V**	**P+S**	**V+S**	**P+V+S**
3	30.7	36.4	40.0	50.7	59.4	58.8
4	30.5	38.5	42.2	53.0	61.2	59.9
5	30.8	40.1	44.9	53.3	62.1	60.4
6	31.3	40.1	46.0	53.0	61.8	60.8
7	31.0	37.2	44.7	53.3	62.4	62.2
8	33.4	40.3	45.3	53.1	63.9	60.2
9	32.5	39.6	44.7	54.7	64.1	61.1
ms 3 to 9	32.5	40.5	43.5	54.4	**64.5**	61.4

Handshape analysis

One of the advantages of KL-HMM approach is that the parameters i.e. the categorical distribution of HMM states can be interpreted. Figure 6.5 shows the handshape categorical distributions of the 9 states **V+S** system of two signs: AUCH and KRANK. In the AUCH case, the **V** system recognized 0 samples out of 9, the **S** system 3 samples and the **V+S** one 6 samples; thus, we can hypothesize that the handshape information is the major source of information in that case. Indeed, the density plot of the handshape categorical distributions shows that the model contains relevant information since the sequence of maximum distribution by state (1, 1, 1, 37, 37, 1, 37, 1, 1) corresponds to the true label (1, 37, 1). In the KRANK case, the reverse can be observed;

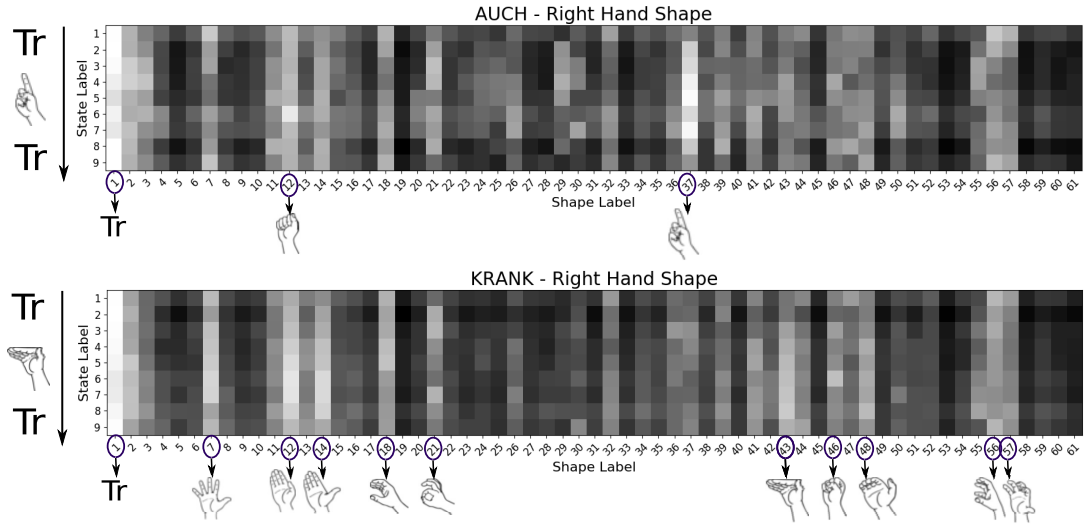

Figure 6.5 – Density plots of the right handshape categorical distribution linked to each KL-HMM states for AUCH and KRANK sign's model. **Tr** is used for the Transition shape.

adding the handshape features adds confusion in the recognition task. The **V** system recognized 5 samples out of the 7, the **S** system 1 sample and **V+S** 3 samples. The density plot confirms the fact that there is a confusion in the model itself since the resulting handshapes are the transition shape for all the states. This can be partly attributed to high signer variations in the handshapes used in the training data.

6.4 Multilingual Sign Language Recognition

In Chapter 5, we observed that the inferred hand movement subunits exhibit language independence characteristics. Building upon that, in this section, we extend the proposed phonology-based framework to multilingual SLR, where the hand movement information is modeled using multiple sign language resources, similar to handshape information.

6.4.1 Proposed multilingual framework

The KL-HMM framework can be visualized as matching of a sequence of multichannel information obtained through bottom-up modeling (visual signal-to-hand gestures) with a sequence of multichannel information obtained through top-down modeling (lexeme-to-hand gestures). In KL-HMM based speech recognition, it has been found that resource constraints can be effectively addressed by using auxiliary or non-target language resources for bottom-up modeling and using the target language resources only for top-down modeling [93]. Given that understanding, a

question that arises is: can we achieve the same for sign language recognition? In other words, $\mathbf{z}_t$ estimators are trained with target language independent data and $\mathbf{y}_{s^i}$ is estimated on target language data. The monolingual study presented earlier demonstrates that capability for modeling handshape.

The subunits approach developed in Chapter 5 paves the path for such a multilingual approach for modeling hand movement. More precisely, as depicted in Figure 6.6, one hand movement subunit estimator is trained for each auxiliary sign language separately to estimate hand movement subunits probabilities. These estimates are then stacked, and KL-HMM parameters are trained on the target language data.

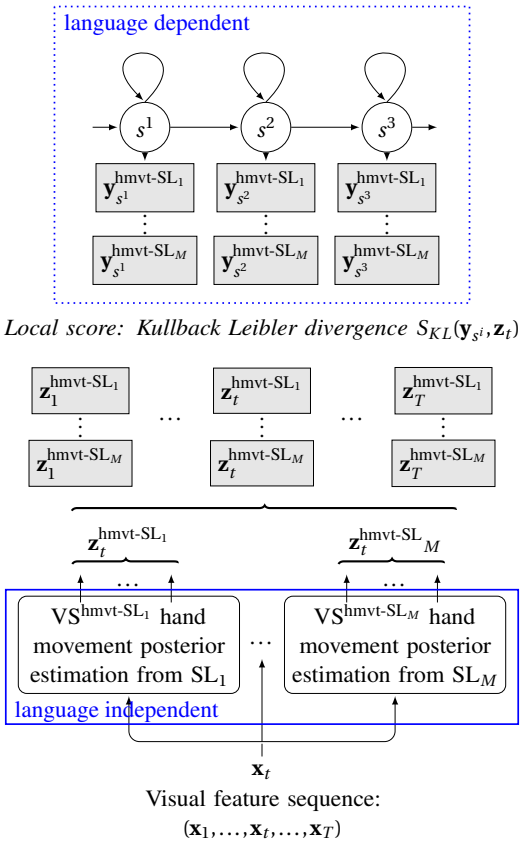

Figure 6.6 – Illustration of the adapted KL-HMM-based phonology-based framework to multilingual scenario, where the hand movement visual subunits (VS$^{\text{hmvt}}$) are extracted from different sign languages (SL$_1$ to SL$_M$).

6.4.2 Experimental setup

To validate the proposed multilingual framework, we derived the language independent hand movement subunits from three different languages and tested them through development of cross- and multi-lingual systems.

Databases

In this experiment, we used databases from three sign languages, namely the Swiss German Sign Language SMILE DSGS database, the Turkish Sign Language HospiSign database and the German Sign Language DGS database. These databases are described in Section 2.3.

For the SMILE DSGS database, we used the same setup as in the monolingual study (see Section 6.3.1). In order to conduct a signer-independent experiment with the HospiSign and DGS databases, we followed leave-one-signer out protocol. In the case of HospiSign database, in each fold on average there were 1074 training samples and 210 test samples. In the DGS database, in each fold on average there were 2586 training samples and 227 test samples.

Handshape subunit estimator

We use the same off-the-shelf DeepHand handshape subunits posterior probability estimator that was used in the monolingual study (see Section 6.3.2).

Hand movement subunit estimator

The hand movement subunit posterior probability estimators were based on the developments made in Chapter 5. Briefly, as illustrated in Figure 6.7, the shoulder normalization-based features described in Section 3.1.1 were used as feature observation, where the skeletons of signers in the SMILE DSGS, HospiSign and DGS corpus were aligned w.r.t a signer from HospiSign database. Then a sign level HMM with one mixture Gaussian and diagonal covariance was trained for each sign using 3 to 30 states. The development data was decoded using all the 28 whole sign-based HMM/GMM for all the signs, and the most frequently recognized model in terms of number of states was chosen. These states served as the hand movement subunits. Then the HMM states was clustered by pairwise comparison of respective Gaussian distributions using the Bhattacharyya distance leading to a clustered subunits states. For building the sign-based and SU-based MLPs, we first obtained the alignments in terms of the HMM states using either the subunits infer by the sign level HMM/GMM (sign-based) or the clustered subunits (SU-based). We then trained MLPs classifying HMM states with output non-linearity of softmax and minimum cross-entropy error criterion. We used the feature observation with four frames preceding context and four

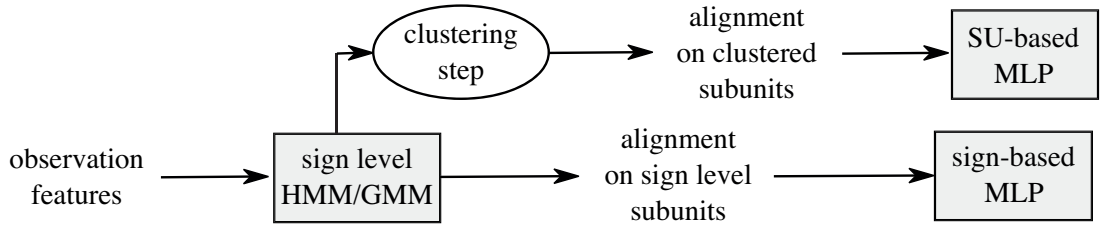

Figure 6.7 – Illustration of the derivation of the SU-based and the sign-based MLP based on the hand movement subunit extraction methods developed in Chapter 5

frames following context as the MLP input. In our experiments, we trained MLPs with different number of hidden units (600, 800, 1000) and hidden layers (0, 1, 2, 3). The number of hidden units and hidden layers as well as other hyper-parameters such as learning rate and the batch size were chosen according to the frame-level accuracy on the development set. For HospiSign and German Sign Language (Deutsch Gebärdensprache) (DGS) databases, the data of one of the signers in the training set was used as development set. The MLPs were trained using the Quicknet software [52].

As depicted in Figure 6.6, the hand movement subunits extraction step was done according to each sign language separately leading to a stack of posterior probabilities $\mathbf{z}_t^{\text{hmvt}} := \left[\mathbf{z}_t^{\text{hmvt-SL}_1}\ \mathbf{z}_t^{\text{hmvt-SL}_2}\ \cdots\ \mathbf{z}_t^{\text{hmvt-SL}_N}\right]^T$. The reason for that is that when we tried to extract a common set of subunits from different corpora, we noticed that during the clustering step the subunits remained separate by languages. This can be explained by the differences in the recording conditions in the different data sets.

Recognition studies

Two studies were conducted: one based on the hand movement subunits solely and a second based on both the hand movement and the handshape subunits. In all cases, we extracted hand movement subunits from either one language (cross-lingual setup) or two languages (multilingual setup). We also developed a corresponding KL-HMM-based monolingual reference as baseline. All the models was trained with 3 to 30 states. For a better visualization, we report the performance of the best system.

6.4.3 Results and analysis

In this section, we first present studies modeling hand movement alone, and then present studies that model both hand movement and handshape.

Hand movement study

Table 6.3 presents the results of the monolingual reference systems and the cross- and multi-lingual systems in terms of recognition accuracy RA (± standard deviation). It can be observed

Table 6.3 – Average RA (± standard deviation), over the leave-one-signer out protocol, for reference monolingual systems and cross-/multi-lingual KL-HMM-based systems using hand movement subunits

(a) Targeted language: DSGS (SMILE DSGS database)

hmvt MLP trained on	KL-HMM			
	sign-based MLP		SU-based MLP	
	dim.	RA	*dim.*	RA
DGS	*281*	46.6	*160*	47.3
TSL	*496*	41.6	*324*	41.5
DGS and TSL	*777*	48.2	*484*	48.4
DSGS	*2257*	57.4	*1946*	55.8

(b) Targeted language: DGS (DGS database)

hmvt MLP trained on	KL-HMM			
	sign-based MLP		SU-based MLP	
	dim.	*RA±std*	*dim.*	*RA±std*
TSL	*496*	52.5 ± 10.2	*324*	52.2 ± 9.5
DSGS	*2163*	57.3 ± 9.8	*1485*	58.1 ± 9.5
TSL and DSGS	*2659*	57.7 ± 9.8	*1809*	58 ± 10.8
DGS	*281*	65.8 ± 13.1	*217*	68.2 ± 10

(c) Targeted language: TSL (HospiSign database)

hmvt MLP trained on	KL-HMM			
	sign-based MLP		SU-based MLP	
	dim.	*RA±std*	*dim.*	*RA±std*
DGS	*281*	97.5 ± 1.4	*160*	95.4 ± 2.0
DSGS	*2163*	98.0 ± 1.1	*1485*	98.8 ± 1.0
DGS and DSGS	*2444*	98.1 ± 1.1	*1645*	98.2 ± 1.1
TSL	*300*	97.5 ± 1.7	*217*	97.3 ± 1.7

that for DSGS (Table 6.3 (a)) and DGS (Table 6.3 (b)) as target languages the performance of cross- and multi-lingual systems are well above random classification but below monolingual system performance. The low performance can be due to combination of two factors: (a) differences in recording settings. More precisely in the SMILE DSGS database the signs are performed sitting while in the DGS and HospiSign databases standing. Skeleton alignment may not fully compensate for these differences. (b) Vocabulary in each database is limited. As a

consequence, not all possible movements can be expected to be covered by the derived subunits. Moreover, the HospiSign database is composed of phrases while the two other databases are composed of isolated signs. These differences can also influence the nature of the extracted subunits. This could explain why adding TSL subunits does not significantly help to recognize DGS or DSGS languages.

Together these results indicate that whether we take subunits from sign-based or SU-based approaches, they exhibit sign language independence characteristics. Indeed, when comparing SU-based MLP and sign-based MLP systems, it can be observed that the performances are comparable, despite the fact that subunit extraction leads to state space reduction of 31% on DSGS, 35% on TSL and 43% on DGS.

Hand movement and handshape study

Table 6.4 presents the results of the handshape based KL-HMM system in terms of recognition accuracy on the three different sign languages (DSGS, DGS and TSL). As it can be observed, for all the three databases, the handshape component is not as good as the hand movement to differentiate the signs. We observed similar trend in the monolingual study. One of the reasons could be the hand orientation independence of the handshape estimator. The cropping of the hand zone is also dependent on the precision of the joint tracking which differs for each database. The very low performance on DGS could be due to the uncontrolled environment in which the database was collected. As the performance is very poor, we decided not to pursue the DGS study modeling both the hand movement and handshape subunits.

Table 6.4 – Average RA (± standard deviation) of the handshape based KL-HMM systems on three sign languages (DSGS, TSL and DGS)

	hshp-based KL-HMM-based	
	dim.	RA
DSGS	*122*	38.2
DGS	*122*	5.8 ± 2.5
TSL	*122*	83.8 ± 8.0

In the next experiment, we combined the hand movement and shape observation to train the systems. Table 6.5 presents the results of the reference monolingual system and cross-/multi-lingual KL-HMM systems. As expected, the handshape component gives complementary information to the hand movement as evidenced by the results. Moreover, adding the handshape decreases the gap in between the monolingual and the cross-/multi-lingual framework. Also, it is interesting to note that the best reported recognition accuracy using both hand movement and handshape information on the HospiSign database is 96.67% (± 1.80) [18].

Table 6.5 – Average RA (± standard deviation) for reference monolingual system and cross-/multi-lingual KL-HMM systems using hand movement and handshape subunits

(a) Targeted language: DSGS (SMILE database)

hmvt MLP trained on	hshp MLP trained on	KL-HMM	
		sign-based MLP	SU-based MLP
DGS	1 million hand	72.9	72.6
TSL	1 million hand	67.3	66.1
DGS and TSL	1 million hand	72.9	73.2
DSGS	1 million hand	75.6	74.3

(b) Targeted language: TSL (HospiSign database)

hmvt MLP trained on	hshp MLP trained on	KL-HMM	
		sign-based MLP	SU-based MLP
DGS	1 million hand	98.6 ± 1.4	99.0 ± 1.1
DSGS	1 million hand	99.0 ± 1.2	99.1 ± 1.1
DGS and DSGS	1 million hand	99.4 ± 0.7	99.3 ± 0.9
TSL	1 million hand	98.9 ± 0.9	98.9 ± 1.4

6.4.4 Multilingual sign language recognition with HMM/GMM approach

The tandem feature-based HMM/GMM framework can also be adapted to multilingual SLR. As depicted in Figure 6.8, the hand movement subunit estimators are trained on different languages and then are transformed using the tandem technique (see Section 3.3.1) before being concatenated and used as feature observation for HMM/GMM system trained in a language-dependent manner. As illustrated in the figure, the target language data is used to estimate both KLT transformation matrix and the parameters of the HMM/GMM system.

Experimental setup

To validate the tandem feature based multilingual framework, we used the setup with DSGS as the target language. The handshape and the hand movement estimators remained the same as in the KL-HMM based multilingual study. We derived the language-independent hand movement subunits from two different languages, namely Turkish Sign Language (TSL) in HospiSign database and German Sign Language (DGS) in DGS database to recognize Swiss German Sign Language (DSGS) isolated signs in the SMILE DSGS database. During tandem feature extraction, in the KLT step, we reduced the feature dimension to cover up to 99% variance of the data. HMM/GMM systems were trained with different number of states, i.e. from 3 to 30 states, similar to the study using KL-HMM. The emission distribution at each state was modeled by one mixture Gaussian distribution (i.e. single multivariate Gaussian distribution).

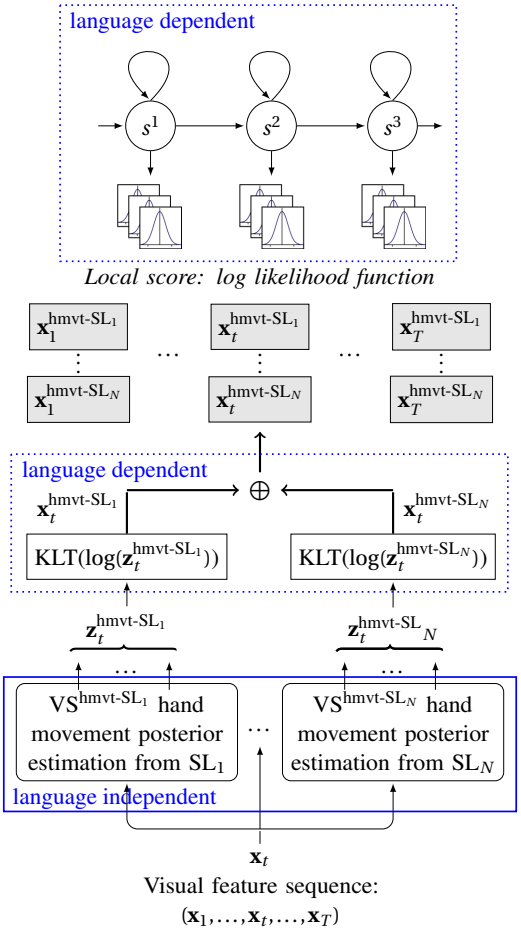

Figure 6.8 – Illustration of the adapted HMM/GMM-based framework to multilingual scenario, where the hand movement visual subunits (VS^{hmvt}) are extracted from different sign languages (SL_1 to SL_M).

Hand movement study

Table 6.6 presents the results of the tandem feature-based cross- and multi-lingual systems in terms of recognition accuracy *RA* for the hand movement study. We can observe that there is a dimensionality reduction of about 96% using the sign-based subunits MLP model and about 91% using the SU-based MLP. The resulting systems lead to higher performance than the KL-HMM approach (Table 6.3 (a)). This indicates that not all language-independent subunits dimensions are of interest for the target language. In the KL-HMM approach, all the language-independent subunits dimensions are modeled, which can be noisy. Furthermore, the KLT matrix is estimated on the target language data, which could compensate for domain differences when extracting tandem features. We can see that particularly for the cross-lingual system using TSL subunits. In

Table 6.6 – RA of the cross-/multi-lingual HMM/GMM-based systems using hand movement subunits

Targeted language: DSGS (SMILE DSGS database)

hmvt MLP trained on	HMM/GMM-based			
	sign-based MLP		SU-based MLP	
	dim.	*RA*	*dim.*	*RA*
DGS	*11*	48.0	*23*	52.6
TSL	*16*	54.1	*20*	54.7
DGS and TSL	*27*	54.6	*43*	56.4

the KL-HMM approach, no such domain difference compensation happens.

Hand movement and handshape study

Table 6.7 presents the results of the handshape based system in terms of recognition accuracy on the SMILE DSGS database. With the criteria of 99% variance, the KLT step reduces the feature dimension by about 60%. The tandem feature based handshape alone system outperforms the handshape alone KL-HMM system (Table 6.4). Interestingly, this performance is better than the monolingual KL-HMM-based hand movement system presented in Table 6.3 (a).

Table 6.7 – RA of the handshape based HMM/GMM system on the SMILE DSGS database

hshp-based HMM/GMM-based

	dim.	*RA*
DSGS	*48*	59.5

Table 6.8 presents the results for cross- and multi-lingual systems when modeling both handshape and hand movement. We observe again that tandem feature based system yields better cross-lingual and multilingual SLR systems than the KL-HMM approach (see Table 6.5 (a)). Furthermore, cross-lingual system and multilingual system performances are similar. This again indicates that domain differences are better compensated by tandem feature-based approach.

Table 6.8 – RA of the cross-/multi-lingual HMM/GMM based systems using hand movement and handshape subunits

Targeted language: DSGS (SMILE DSGS database)

hmvt MLP trained on	hshp MLP trained on	HMM/GMM-based	
		sign-based MLP	SU-based MLP
DGS	1 million hand	77.1	77.8
TSL	1 million hand	77.8	78.3
DGS and TSL	1 million hand	78.0	77.3

6.5 Summary

The focus of this chapter was on developing approaches to jointly model the multichannel information inherent in sign languages. In that context, we argued and showed that the methods developed for articulatory feature modeling in speech processing can be adopted to jointly model the multichannel information for sign language processing. We studied two approaches: a HMM/GMM approach and a KL-HMM approach. Through monolingual and cross-lingual sign language recognition studies, we showed that both approaches succeed in integrating the handshape and the hand movement channel yielding better sign language recognition performance. In Chapter 5, we already demonstrated through a cross-lingual study that the hand movement subunits exhibit sign language sharability property. The cross-/multi-lingual studies in this chapter clearly show that the derived hand movement subunits are transferable across sign languages. These findings are promising, as they pave the path for development of sign language processing systems by sharing multiple sign language resources. We also demonstrated two advantages of the KL-HMM approaches: the decomposition of the hand movement channel and the interpretability of the categorical distribution. The interpretability property of the KL-HMM is what differentiates our framework to other approaches. Indeed, the categorical distribution allows to have a feature space and a time segmentation, i.e. we have access to the production information of each channel separately for each time frame/state. In Chapter 5 (Section 5.3.2 and Section 5.4.2), we somewhat demonstrated this aspect in the context of analysis of hand movement subunits through synthesis of hand movement in the 3D skeleton space. In existing 2-stages approaches [124, 20, 117, 98, 118] which first extract the visual subunit separately and then use them to classify the sign, this property is not obvious. In the following chapter, we build on that to develop an explainable sign language assessment approach.

7 Phonology-based Sign Language Assessment Framework

RQ5: How to assess isolated sign productions at the lexeme-level and the form-level?

Interactive learning platforms are in the top choices to acquire new languages. Such applications or platforms are more easily available for spoken languages, but rarely for sign languages. Assessment of the production of signs is a challenging problem because of the multichannel aspect (e.g., handshape, hand movement, mouthing, facial expression) inherent in sign languages.

In the previous chapter, a HMM-based sign language processing framework was proposed that enables modeling of the multichannel information present in sign languages, akin to modeling of multichannel articulatory information in speech production [94]. The present chapter builds upon that work to propose a sign language assessment approach that, in an integrated manner, can assess sign production at: (a) lexeme level, i.e. verify whether a produced sign is targeting the correct reference sign or not and (b) form level, i.e. assessing separately the different form channels of a sign, such as hand movement and handshape channel. We demonstrate the potential of the proposed approach through a validation study on Swiss German Sign Language.

The chapter is organized as follows: in Section 7.1, we present the related work. Our proposed phonology-based sign language assessment approach is explained in Section 7.2. Section 7.3 and Section 7.4 present the experimental setup and the results and analysis, respectively. Analyses are presented in Section 7.5, Section 7.6 and Section 7.7. Section 7.8 presents an overview of the demonstrator and finally the summary is given in Section 7.9. The basic approach was published in the following publication:

A phonology-based approach for isolated sign production assessment in sign language, Sandrine Tornay, Necati Cihan Camgoz, Richard Bowden and Mathew Magimai.-Doss, in: Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion), 2020

7.1 Related Work

In recent years, there is growing interest in developing assistive systems that can help in bridging the gap or breaking the barrier between hearing and Deaf communities through multimodal systems. In that direction, as sign languages are under-studied and under-resourced languages, there is interest in developing interactive applications that could aid in sign language acquisition. Currently, the existing platforms test comprehension and vocabulary through pre-recorded videos, while sign language production tests are realized by online recording for later analysis, which is both expensive and time consuming. Existing interactive e-learning platforms that contain production testing use either self-correctness, such as the web-based e-learning resource SignAssess [23] which allows to compare the recorded user's video to a pre-recorded reference one, or real-time sign language verification which assesses if the produced sign is correct or incorrect, such as SignAll [112] technology or ISARA [49] application. Assessing whether a produced sign is correct or incorrect would not be sufficient by itself to aid sign language learners. The reason being that sign language consists of different channels of information corresponding to manual components (hand position, hand movement and handshape) and non-manual components (mouthing, facial gesture, posture). So, for realistic adoption of sign language learning applications, there is need for a framework that enables assessment of those multiple channels of information in a linguistically valid manner.

7.2 Proposed Phonology-based Framework for Sign Language Assessment

In Chapter 6, a phonological approach for SLR was presented, based on the understanding that, both sign language and spoken language are communication process that imply: a source, i.e. a production phenomenon that generates a signal and a receiver, i.e. a perception phenomenon which interprets the generated signals in terms of elements of "language", e.g. words, phrases. Based on this understanding, we proposed the KL-HMM approach for sign language recognition. At a high level, as can be seen in Figure 6.4, the KL-HMM approach can be visualized as an approach where,

1. the element of sign language in the perception space modeled through HMM states is projected onto production space, yielding a "reference" sequence of stacked hand movement and handshape subunits posterior probability sequence;

2. the input visual signal is projected onto the production space, yielding a "test" sequence of sequence stacked hand movement and handshape subunits posterior probability sequence; and

3. matching the reference and test sequences of posterior probability sequences through dynamic programming (Viterbi algorithm) to determine the match between the element of language modeled through the HMM and the observed visual signal resulting from sign production.

The KL-HMM approach can be as it is adopted for sign language assessment by adding a verification or decision-making step on top of the matching process.

More precisely, in the KL-HMM framework, we can cast the sign language assessment problem as matching a test sign production (test sequence of probabilities) against an "expected" sign production (reference sequence of probabilities) and deciding whether the test sign production is acceptable or not. The decision-making can be carried out in a relatively easy manner, as comparison of probability distributions using KL-divergence and other measures such as Bhattacharya distance is equivalent to hypothesis testing [12, 55]. In other words, by simply applying a threshold on the resulting KL-divergence based matching score the decision about acceptability of test sign production can be made. We will see later in this section that the threshold can be applied at different levels such as, at the lexeme level, at the individual channel level. Thus, leading to an explainable sign language assessment approach that can carry out the assessment at different levels in an integrated manner.

In principle, besides KL-HMM, the reference/expected sequence of subunits posterior probabilities can be also obtained in an instance-based manner. More precisely, given the visual signal of an "acceptable" production of a sign, the expected sequence of subunits posterior probabilities for that sign can be obtained by feeding the visual signal as input to the different subunits posterior probability estimators, as done for obtaining the test sequence of subunits posterior probabilities from the test sign production.

To make a distinction between the two methods to obtain the expected sequence of subunits posterior probabilities, we refer to the KL-HMM based approach as "multiple views based reference", as the KL-HMM is trained on multiple signers data. While, we refer to the instance-based approach as "single view based reference", as a single acceptable sign production by a signer is used to obtain the reference sequence.

7.2.1 Multiple views reference method

As explained earlier, there are two processes involved in sign language assessment: (a) matching process and (b) decision-making process.

Matching process

As illustrated in Figure 7.1, given the expected sign, the KL-HMM generates the reference sequence of stacked categorical distributions $(\mathbf{y}_1, \cdots, \mathbf{y}_N)$, which is matched with test sequence of stacked probability distributions $(\mathbf{z}_1, \cdots, \mathbf{z}_T)$ obtained for test sign production using the subunits posterior probability estimators. Formally, the match is obtained by dynamic programming

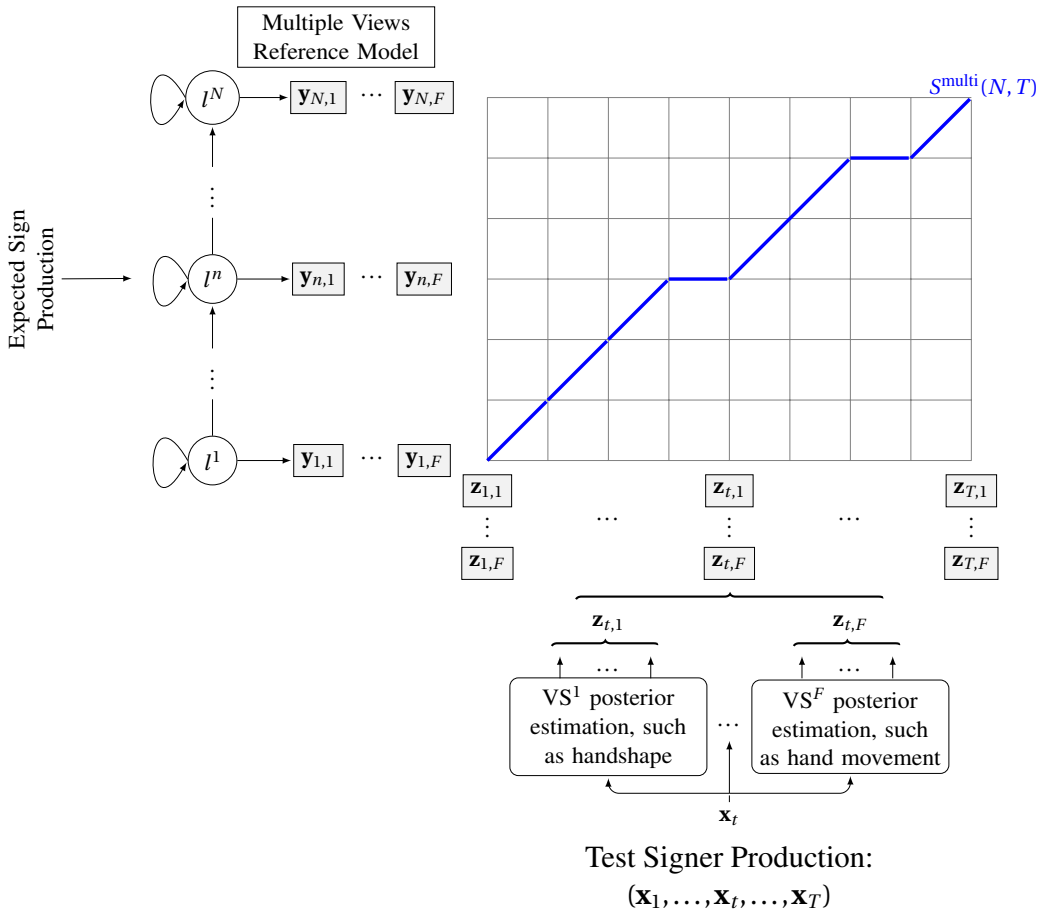

Figure 7.1 – Illustration of the phonology-based assessment framework using the multiple views reference model.

(Viterbi algorithm) with the following recursion,

$$S^{\text{multi}}(n,t) = l(\mathbf{y}_n, \mathbf{z}_t) + \min\left[S^{\text{multi}}(n,t-1) + c_{\text{tr}}, S^{\text{multi}}(n-1,t-1) + c_{\text{tr}}\right], \tag{7.1}$$

where $c_{\text{tr}} = -\log(0.5)$ is the transition cost and $l(\cdot,\cdot)$ is the local score defined by the Symmetric Kullback-Leibler (SKL)-divergence between two probability distributions, i.e.,

$$l(\mathbf{y}_n, \mathbf{z}_t) = \sum_{f=1}^{F} S_{SKL}(\mathbf{y}_{n,f}, \mathbf{z}_{t,f}), \tag{7.2}$$

where

$$S_{SKL}(\mathbf{y}_{n,f}, \mathbf{z}_{t,f}) = \frac{1}{2} \cdot \sum_{d=1}^{D_f} \left(\mathbf{y}_{n,f}^d \log(\frac{\mathbf{y}_{n,f}^d}{\mathbf{z}_{t,f}^d}) + \mathbf{z}_{t,f}^d \log(\frac{\mathbf{z}_{t,f}^d}{\mathbf{y}_{n,f}^d})\right); \tag{7.3}$$

where $\mathbf{y}_{n,f}^d$ and $\mathbf{z}_{t,f}^d$ denote d^{th} element in the vectors $\mathbf{y}_{n,f}$ and $\mathbf{z}_{t,f}$, respectively, $f \in \{1, \cdots F\}$ denotes a channel and F denotes the number of channels. The best matching path in the reference lexeme can be obtained as part of the dynamic programming recursion.

Decision-making process

Given the best matching path, a lexeme-level and a form-level scores, $\mathscr{S}_{lex}^{\text{multi}_1}, \mathscr{S}_{form,f}^{\text{multi}_1}$ respectively, are estimated by

$$\mathscr{S}_{lex}^{\text{multi}_1} = \frac{1}{T} \cdot \sum_{n=1}^{N} \sum_{t=t_n^b}^{t_n^e} l(\mathbf{y}_n, \mathbf{z}_t), \tag{7.4}$$

and by

$$\mathscr{S}_{form,f}^{\text{multi}_1} = \frac{1}{T} \cdot \sum_{n=1}^{N} \sum_{t=t_n^b}^{t_n^e} S_{SKL}(\mathbf{y}_{n,f}, \mathbf{z}_{t,f}), \tag{7.5}$$

where t_n^b and t_n^e are the begin/end time frames, respectively of each state n, of the best matching path; T is total number of frame; and the form-level score assesses the channel f.

Lexeme-level and form-level assessment can be carried out by simply applying a threshold on $\mathscr{S}_{lex}^{\text{multi}}, \mathscr{S}_{form,f}^{\text{multi}}$ to decide correct/incorrect lexeme and forms.

The decision can also be taken by normalizing the state duration information. More precisely,

$$\mathscr{S}_{lex}^{\text{multi}_2} = \frac{1}{N} \cdot \sum_{n=1}^{N} \frac{\sum_{t=t_n^b}^{t_n^e} l(\mathbf{y}_n, \mathbf{z}_t)}{t_n^e - t_n^b + 1}, \tag{7.6}$$

and

$$\mathscr{S}^{\text{multi}_2}_{form,f} = \frac{1}{N} \cdot \sum_{n=1}^{N} \frac{\sum_{t=t_n^b}^{t_n^e} S_{SKL}(\mathbf{y}_{n,f}, \mathbf{z}_{t,f})}{t_n^e - t_n^b + 1} \,. \tag{7.7}$$

We refer to the first one as frame level normalization and the second one as state level normalization.

7.2.2 Single view reference method

Similar to the multiple views reference method, there is a matching process followed by a decision-making process.

Matching process

As illustrated in Figure 7.2, the methodology remains the same as in the multiple views reference method, where the sequence of stacked categorical distributions of KL-HMM states $(\mathbf{y}_1, \cdots, \mathbf{y}_N)$ of the reference model are replaced by the sequence of stacked probability distributions $(\mathbf{z}_1^{\text{ref}}, \cdots, \mathbf{z}_{T'}^{\text{ref}})$ obtained from an acceptable production of the expected sign. Formally, the match is obtained by dynamic programming with the following recursion,

$$S^{\text{spl}}(t', t) = l(\mathbf{z}_{t'}^{\text{ref}}, \mathbf{z}_t) + \min\left[S^{\text{spl}}(t', t-1), S^{\text{spl}}(t'-1, t), S^{\text{spl}}(t'-1, t-1)\right] , \tag{7.8}$$

where t' denotes the time frame of the reference sequence and $l(\cdot,\cdot)$ is the local score obtained by computing SKL-divergence (see Equation (7.2)). The matching process also yields the best matching path.

Decision-making process

Given the best matching path, the lexeme-level score and form-level score can be obtained like frame level normalization in the multiple views reference method. More precisely, the lexeme-level score $\mathscr{S}^{\text{spl}}_{lex}$ is computed by summing the local scores $l(\cdot,\cdot)$ on the best matching path and normalizing it by the path length. The form-level score is obtained by summing the channel specific SKL-divergence score on the best matching path and normalizing it by the path length.

7.3 Experimental Setup

We validated the proposed approach on the linguistically annotated SMILE DSGS database. We demonstrate the approach using two channels of information for which annotations are available,

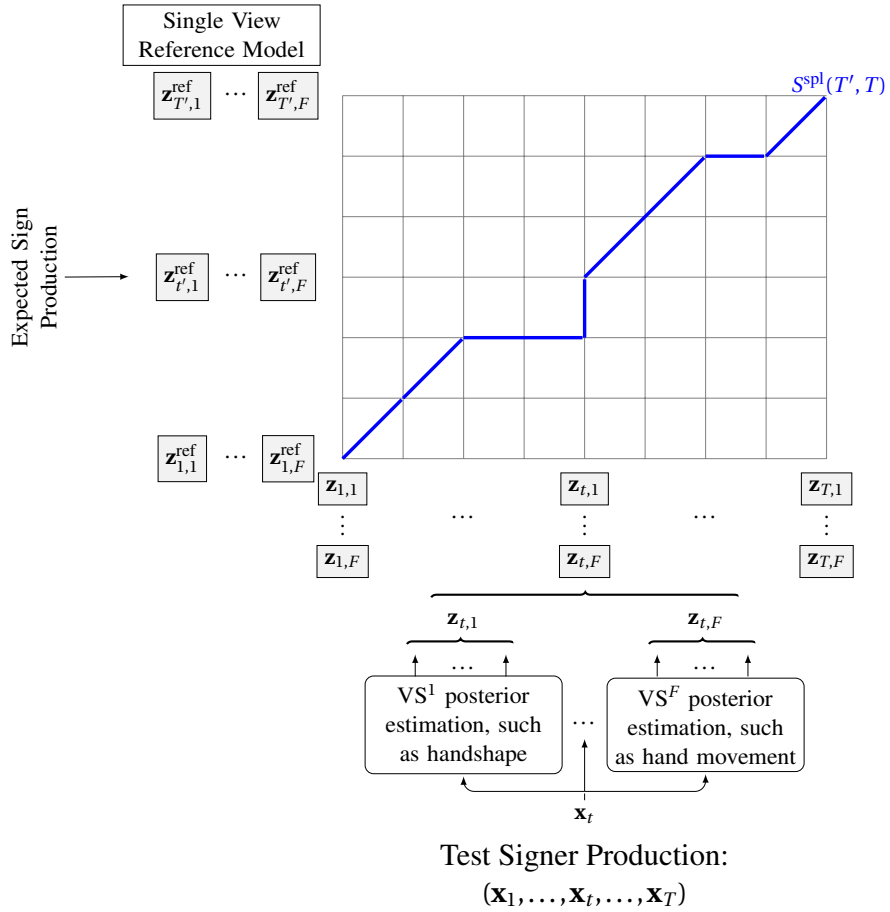

Figure 7.2 – Illustration of the phonology-based assessment framework using the single view reference model.

namely, hand movement (hmvt) and handshape (hshp).

7.3.1 Database

As described in Section 2.3, the SMILE DSGS database was created in the context of developing an assessment system for lexical signs of Swiss German Sign Language; thus various annotations/transcriptions is available in this database. In our experimental setup, we only used the data annotated with the 'Category of sign produced' annotation of the SMILE transcription/annotation scheme presented in [34]. Briefly, this linguistic annotation evaluates, through six categories, the acceptability of a sign according to linguistic criteria (lexeme, meaning and form), see Table 7.1. The cat.1 and cat.2, consisting of acceptable sign productions, was partitioned in a signer-independent manner into 1125 training set samples from 15 signers, 509 development

set samples from 7 signers and 581 test set samples from 8 signers. We used the same test set samples for evaluating both the KL-HMM model-based references (in terms of SLR) and the proposed sign language assessment systems. The number of test samples of other categories are given in Table 7.1. The cat.1 and cat.2 were used to build the different components of the proposed assessment systems.

Table 7.1 – SMILE annotation scheme of the 'Category of sign produced' annotation

Category	**Same lexeme as target sign?**	**Same meaning as target sign?**	**Same form as target sign?**	*#test samples*
cat.1	yes	yes	yes	581
cat.2	yes	yes	slightly different	
cat.3	yes	yes	no	412
cat.4	yes	slightly different	slightly different	
cat.5	no	yes	no	183
cat.6	no	no	no	

7.3.2 Handshape subunit posterior probability estimation

The handshape feature extraction for sign language assessment presented in Section 3.1.2 was used in this experiment, where the 30-dimensional vector of both hands and the 31-dimensional (with transition shape) of both hands are stacked resulting to a 122-dimensional vector.

7.3.3 Hand movement subunits posterior probability estimation

The hand movement estimator used in this study is the sign-based MLP, illustrated in Figure 6.7, which is built on the sign level hand movements subunits developed in Chapter 5. The feature observations are the shoulder normalization-based features described in Section 3.1.1.

7.3.4 Sign reference systems

We used five compositions of the production space to develop the reference models, namely,

- the **rlS** system refers to the case where only the handshape subunit posterior probabilities of the right and left hands estimated by the residual network-based CNN are stacked and modeled.
- the **M** system refers to the case where only the posterior probabilities of hand movement subunits obtained by combining right and left hand features are modeled. In other words,

distinction between dominant hand and non-dominant hand is not made.

- the **rlM** system refers to the case where hand movement subunits are obtained for the left hand and the right hand separately, i.e. two separated hand movement MLPs classifier are trained. Then the left and right hand movement subunits posterior probabilities estimated by the respective MLPs are stacked and modeled.
- the **rlS+M** and **rlS+rlM** systems refer to the case of using the concatenation of the handshape and the hand movement subunit probability posteriors depending on the different setups presented above.

In the single view reference model, a reference sample was randomly chosen for each sign in the cat.1 data if available otherwise in cat.2; cat.1 and cat.2 correspond to "acceptable sign production" (see Table 7.1).

In the multiple views reference model, we trained five KL-HMM systems corresponding to each production space. Data of the cat.1 and cat.2 (see Table 7.1) were used to train and test the reference models. All the KL-HMM systems were trained using 3 to 30 HMM states per sign. The system that yielded the best recognition accuracy on the development data was chosen as the reference.

Evaluation of the KL-HMM-based reference models: As the experiment setup differs from the SLR studies presented earlier on the SMILE DSGS database, we conducted a sign language recognition study on the KL-HMM reference model. Table 7.2 presents the RA of the different KL-HMM systems. It can be observed that the system modeling both hand movement and handshape information yields the best SLR performance. These results show that the KL-HMM reference lexeme models are indeed modeling the different signs and are able to discriminate between them.

Table 7.2 – Sign language recognition accuracy (RA) of the KL-HMM systems used as the multiple views reference models.

	KL-HMM References				
	rlS	**M**	**rlM**	**rlS+M**	**rlS+rlM**
RA	37.2	56.9	57.4	74.7	75.2

7.3.5 Assessment systems

Lexeme assessment: to evaluate the lexeme assessment, according to the category annotation of the data summarized in Table 7.1, we separated the test correct/incorrect data as the following: *cat.1-2-3-4* which is correct target signs composed of cat.1 to cat.4 and *cat.5-6+* which is incorrect

target signs composed of cat.5, cat.6 and since these categories contain only few data, we balanced the incorrect set by creating additional data by matching each sample of the cat.1 and cat.2 data with a randomly chosen wrong reference.

Form assessment: to evaluate the form assessment, we used the *cat.1-2* as correctly produced form data and since the targeted sign is incorrect for *cat.5-6+* we supposed that the produced form (hand movement and handshape) was incorrect. In the present study, we did not make any difference between dominant and non-dominant hand.

We determined the thresholds, $\delta_{lex}^{\text{spl}}$, $\delta_{lex}^{\text{multi}}$ and $\delta_{form,f}^{\text{spl}}$, $\delta_{form,f}^{\text{multi}}$ for $f \in \{\text{hmvt, hshp}\}$ on the development set, which consists of cat.1 and cat.2 data. More precisely, we created a set of correct sign scores by matching the same sign instances and a set of incorrect match scores by matching instances of different signs. $\delta_{lex}^{\text{spl}}$, $\delta_{lex}^{\text{multi}}$ and $\delta_{form,f}^{\text{spl}}$, $\delta_{form,f}^{\text{multi}}$ for each f were set as the threshold that yielded the best F_1 score for lexeme assessment and form assessment, respectively.

The evaluation measure used in this study was the F_1 score described in Section 2.4.

7.4 Results and Analysis

Lexeme assessment: Table 7.3 presents the F_1 score of the lexeme assessment study using the single view or the multiple views reference models depending on the production space used to align the produced sign. As it can be observed, the multiple views reference model

Table 7.3 – F_1 scores of the correct lexeme assessment using the single view or the multiple views reference models according to the five production space setups

		Production Space				
Reference Model	**Normalization**	**rlS**	**M**	**rlM**	**rlS+M**	**rlS+rlM**
Single view	-	0.73	0.84	0.79	0.83	0.80
Multiple views	frame	0.73	0.88	0.84	0.88	0.84
	state	0.72	0.88	0.85	0.90	0.87

methods outperform the single view reference model methods. A potential explanation is that the KL-HMM models the acceptable variation of the sign, while the single view reference model reference only contain one representation of the sign. Moreover, in the multiple views cases, combining the hand movement and the handshape channel helps since using **rlS+M** as reference gives the best assessment result, while it is the hand movement channel in the single view case. According to the two proposed normalizations, we can observe that there is no statistical difference in the model using one modality (**rlS**, **M** and **rlM**). In the two-modalities models, the state-level normalization gives better lexeme-level assessment than the frame-level. Another relevant observation is that using combined right and left hand movement (**M**, **rlS+M**) is sufficient

for lexeme assessment in both reference cases.

Form assessment: Table 7.4 presents the F_1 score of the forms error assessment study of the hand movement channel and the handshape channel using the single view or the multiple views reference models depending on the five production space setups. First, the same observation as the

Table 7.4 – F_1 scores of the forms error assessment (hand movement (hmvt) and handshape (hshp)) using the single view or the multiple views reference models according to the five production space setups

			Production Space				
Reference Model	**Normalization**	**Form**	**rlS**	**M**	**rlM**	**rlS+M**	**rlS+rlM**
Single view	-	hshp	0.74	-	-	0.76	0.76
		hmvt	-	0.87	0.83	0.85	0.83
Multiple views	frame	hshp	0.77	-	-	0.80	0.80
		hmvt	-	0.88	0.85	0.88	0.87
	state	hshp	0.77	-	-	0.83	0.82
		hmvt	-	0.90	0.86	0.90	0.87

lexeme-level assessment can be made: the multiple views reference model methods outperform the single view methods. Thus, indicating that the acceptable production variation of sign modeled in the KL-HMM is helping the form-level assessment. Moreover, in all the cases, we can observe that adding the hand movement information helps in the handshape error assessment, while the reverse is not true. Indeed, using either **rlS+M** or **rlS+rlM** does not change significantly and is better than using **rlS** for handshape form error assessment. A potential reason for that could be that the hand movement channel has more temporal variations than the handshape channel. This can also explain why adding handshape channel to hand movement one does not help in hand movement error assessment. In fact, hand movement form assessment using **M** or **rlS+M**, or using **rlM** or **rlS+rlM** are not significantly different. Furthermore, making no distinction between dominant and non-dominant hand movement gives better form assessment results. This aspect could be further explained or understood by separating the one-handed or two-handed sign assessment results. Concerning the proposed normalizations, we can observe that the form is better assessed using the state-level normalization.

As the state-level normalization yields better assessment, in the remainder of this chapter, we used the lexeme-level and the form-level scores estimated based on the state-level normalization in the multiple views reference method.

7.5 Impact of Clustered HMM States based Hand Movement Subunits

In the assessment studies presented until now, the hand movement subunits are obtained from sign-level HMM. However, as seen in Chapter 5, the sign-level HMM states can be clustered to reduce the state space and obtain a different set of hand movement subunits. In the SLR studies, we found that both methods yield comparable systems. A question that arises is that: does the same holds for assessment?. We investigated that aspect by using the SU-based MLP estimator (depicted in Figure 6.7) built on the clustered hand movement subunits developed in Chapter 5. We denote,

- **M**$^{\text{SU}}$ to refer to the case where only the clustered hand movement subunits obtained by combining dominant and non-dominant hand features are modeled;
- **rlM**$^{\text{SU}}$ to refer to the case where the clustered hand movement subunits are obtained for the dominant hand and the non-dominant hand separately;
- **rlS+M**$^{\text{SU}}$ and **rlS+rlM**$^{\text{SU}}$ to refer to the case of using the concatenation of the handshape and the clustered hand movement subunit probability posteriors depending on the different setups presented above.

The remainder of the experimental setup is the same as the main study. For easiness of comparison, in the results reported below we also provide the main study results. Table 7.5 presents the SLR studies for the KL-HMM system.

Table 7.5 – Sign language RA of the KL-HMM-based references using the clustered hand movement subunits derivation proposed in Chapter 5 with the dimension of the features (*# feature*)

	KL-HMM-based References - SLR							
	M$^{\text{SU}}$	**M**	**rlM**$^{\text{SU}}$	**rlM**	**rlS+M**$^{\text{SU}}$	**rlS+M**	**rlS+rlM**$^{\text{SU}}$	**rlS+rlM**
RA	51.3	56.9	58.6	57.4	68.0	74.7	74.3	75.2
# feature	*1577*	*2075*	*3503*	*4214*	*122+1577*	*122+2075*	*122+3503*	*122+4214*

Lexeme assessment: Table 7.6 presents the F_1 score of the lexeme assessment study using the single view or the multiple views reference models according to the production space based on the clustered hand movement subunit derivation proposed in Chapter 5. It can be observed in both, the single view or the multiple views reference model, cases that the clustered hand movement subunits helps in the case where the dominant and non-dominant hand movement subunits are computed separately (**rlM**, **rlM+rlS**). It is interesting to note that the same trend can be observed in the recognition accuracies of Table 7.5. But, on the whole, we can observe that the clustered

Table 7.6 – F_1 score of the correct lexeme assessment using the single view or the multiple views reference models according to the production space using the clustered hand movement subunits estimator (SU-based MLP) proposed in Chapter 5 ($\mathbf{M}^{SU}$, $\mathbf{rlM}^{SU}$)

	Production Space							
Reference Model	**M^{SU}**	**M**	**rlM^{SU}**	**rlM**	**rlS+M^{SU}**	**rlS+M**	**rlS+rlM^{SU}**	**rlS+rlM**
Single view	0.82	0.84	0.81	0.79	0.79	0.83	0.81	0.80
Multiple views	0.87	0.88	0.87	0.85	0.90	0.90	0.88	0.87

subunits carry sufficiently enough information for lexeme assessment, since there is no large variation in the lexeme assessment results.

Form assessment: Table 7.7 presents the F_1 score of the form assessment study using the single view or the multiple views reference models according to the production space based on the clustered hand movement subunit derivation proposed in Chapter 5. We can draw the same conclusion as the lexeme assessment, i.e. separation of dominant and non-dominant hands helps and the subunits keep the necessary information on the production of the sign, even with a subunit reduction of 24% in the $\mathbf{M}^{SU}$ case and around 17% in the $\mathbf{rlM}^{SU}$ case.

Table 7.7 – F_1 score of the forms error assessment using the single view or the multiple views reference models according to the production space using the clustered hand movement subunits derivation proposed in Chapter 5 ($\mathbf{M}^{SU}$, $\mathbf{rlM}^{SU}$)

		Production Space							
Reference Model	**Form**	**M^{SU}**	**M**	**rlM^{SU}**	**rlM**	**rlS+M^{SU}**	**rlS+M**	**rlS+rlM^{SU}**	**rlS+M**
Single view	hshp	-	-	-	-	0.77	0.76	0.76	0.76
	hmvt	0.82	0.87	0.84	0.83	0.82	0.85	0.83	0.83
Multiple views	hshp	-	-	-	-	0.82	0.83	0.82	0.82
	hmvt	0.89	0.90	0.88	0.86	0.88	0.90	0.87	0.87

7.6 Impact of Model Selection-based HMM Topology Inference

In Chapter 4, we presented a data-driven HMM-based approach to infer the appropriate number of states dynamically during the recognition. In the studies presented until now, we have used a fixed common number of states for all the signs based on the best recognition accuracy on the development data. We conducted a study with this model selection approach. We denote,

- the $\mathbf{rlS}^{\#}$ to refer to the case where only the handshape subunit posterior probabilities of the dominant and non-dominant hands estimated by the residual network based CNN are modeled.
- $\mathbf{M}^{\#}$ to refer to the case where only the sign-level hand movement subunits obtained by

combining dominant and non-dominant hand features are modeled;

- **rlM**$^{\#}$ to refer to the case where the sign-level hand movement subunits are obtained for the dominant hand and the non-dominant hand separately;

- **rlS+M**$^{\#}$ and **rlS+rlM**$^{\#}$ to refer to the case of using the concatenation of the handshape and the hand movement subunit probability posteriors depending on the different setups presented above.

The same experimental setup as in the main study was used in this analysis.

SLR accuracies: Table 7.8 presents the sign language RA of the KL-HMM-based references using the data-driven HMM-based structure derivation proposed in Chapter 4 with the corresponding mean of the number of states with the corresponding number of states. It can be observed, the

Table 7.8 – Sign language RA of the reference KL-HMM systems using the data-driven HMM-based structure derivation proposed in Chapter 4 with the corresponding mean of the number of states (*mean / # state*)

	KL-HMM References									
	rlS$^{\#}$	**rlS**	**M**$^{\#}$	**M**	**rlM**$^{\#}$	**rlM**	**rlS+M**$^{\#}$	**rlS+M**	**rlS+rlM**$^{\#}$	**rlS+rlM**
RA	35.1	37.2	56.5	56.9	55.9	57.4	73.3	74.7	73.8	75.2
mean / # state	17	26	26	18	26	24	17	29	17	28

model selection based HMM structure derivation does not gains over the fixed HMM topology obtained based on the recognition accuracy on the development set in the SLR framework. It is interesting to notice that the mean of the number of states is lower than the fixed number of states derived from the development data in all the KL-HMM-based references containing the handshape information while in the hand movement KL-HMM-based references case it is upper. But as depicted in Figure 7.3, in the case of the **rlS+rlM**$^{\#}$ model, the most chosen number of states is 26 and the fixed number of states of the **rlS+rlM** model is 28.

Table 7.9 and Table 7.10 present lexeme assessment and form assessment results, respectively. Although there is slight drop in SLR accuracy, the model selection approach does not really affect lexeme assessment and form assessment.

Table 7.9 – F_1 score of the correct lexeme assessment using the KL-HMM-based reference adapted with the data-driven HMM-based structure derivation proposed in Chapter 4

KL-HMM References									
rlS$^{\#}$	**rlS**	**M**$^{\#}$	**M**	**rlM**$^{\#}$	**rlM**	**rlS+M**$^{\#}$	**rlS+M**	**rlS+rlM**$^{\#}$	**rlS+rlM**
0.74	0.72	0.88	0.88	0.86	0.85	0.90	0.90	0.87	0.87

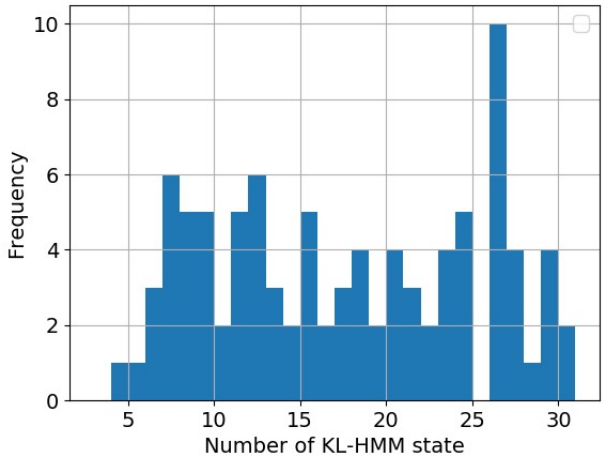

Figure 7.3 – Frequency of the selected number of states of the **rlS+rlM**$^{\#}$ model.

Table 7.10 – F_1 score of the forms error assessment using the KL-HMM-based references adapted with the data-driven HMM-based structure derivation proposed in Chapter 4

	KL-HMM References									
Form	**rlS**$^{\#}$	**rlS**	**M**$^{\#}$	**M**	**rlM**$^{\#}$	**rlM**	**rlS+M**$^{\#}$	**rlS+M**	**rlS+rlM**$^{\#}$	**rlS+rlM**
hshp	0.77	0.77	-	-	-	-	0.83	0.83	0.82	0.82
hmvt	-	-	0.90	0.90	0.86	0.86	0.88	0.90	0.88	0.87

7.7 Interpretable Assessment Score

Until now, we have used the lexeme-level score and form-level score obtained based on SKL-divergence for decision-making. While this is sufficient from assessment perspective, an interpretable score is desirable for providing feedback. For instance, in scenarios like sign language learning, besides providing the final decisions of lexeme-level and form-level assessments, it would be good to provide a score that indicates the "confidence" with which the decision was taken. In this section, we show that the SKL-based lexeme-level score and form-level score can be converted into a posterior-based confidence measure. As KL-divergence yields an estimate of log-likelihood ratio [12]. One way to obtain posterior-based confidence measure Conf is,

$$\text{Conf} = \frac{2}{1 + \exp(\mathscr{S})}, \tag{7.9}$$

where $\mathscr{S} \in \{ \mathscr{S}^{\text{multi}}_{lex}, \mathscr{S}^{\text{spl}}_{lex}, \mathscr{S}^{\text{spl}}_{form,f}, \mathscr{S}^{\text{multi}}_{form,f} \}$. As KL-divergence varies between 0 and $+\infty$, Conf varies between 1 and 0.

Figures 7.4 compares the histogram of the SKL scores $\mathscr{S}^{\text{multi}}_{lex}$ and the derived Conf score of the **rlM+rlS** model according to the cat.1-2, cat.3-4 and cat.5-6o data, the corresponding threshold is

drawn as a dashed line. Firstly, we can notice that the SKL scores of the correct lexeme production

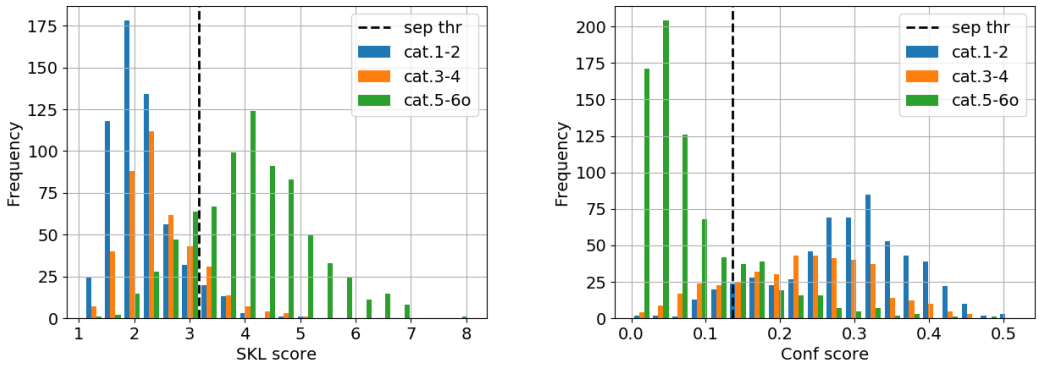

Figure 7.4 – Histograms of the SKL scores $\mathscr{S}_{lex}^{\text{multi}}$ and the derived Conf scores of the **rlM+rlS** model.

of sign (cat.1-2-3-4) and the incorrect (cat.5-6o) are separable and that the method used to set the threshold is adapted to the task. Secondly, same inferences can be done at lexeme-level and form-level, if we apply a threshold on the posterior-based confidence measure. We can see on the Conf score histogram that the Conf scores never reaches 1.0. In other words, we obtain an under estimate. This is due to the fact that a perfect match i.e. KL-divergence equal to 0 is quite improbable, since we are comparing probability distributions. One way to address that issue is to add an offset value α when computing the confidence score as follows

$$\text{Conf}_{\alpha} = \frac{2}{1 + \exp\left(\mathscr{S} - \alpha\right)} \,. \tag{7.10}$$

In the present experiment, we set α as the minimum SKL score obtained on the cat.1-2 development data. Figure 7.5 compares the histogram of the Conf and the Conf_{α} scores of the **rlM+rlS** model. As depicted by the figure, the offset value α has the expected effect.

We carried out lexeme assessment and form assessment studies using Conf_{α}. We denote the respective models using the posterior-based confidence measure Conf_{α} as,

- the $\textbf{rlS}_{\text{conf}}$ to refer to the case where only the handshape subunit posterior probabilities of the dominant and non-dominant hands estimated by the residual network based CNN are modeled.
- $\textbf{M}_{\text{conf}}$ to refer to the case where only the sign-level hand movement subunits obtained by combining dominant and non-dominant hand features are modeled;
- $\textbf{rlM}_{\text{conf}}$ to refer to the case where the sign-level hand movement subunits are obtained for the dominant hand and the non-dominant hand separately;

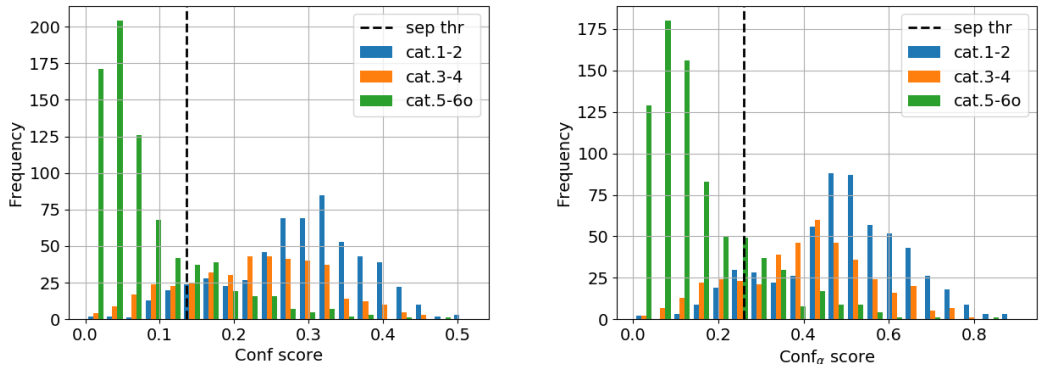

Figure 7.5 – Histograms of the Conf$_\alpha$ scores and the Conf scores of the **rlM+rlS** model.

- **rlS+M**$_{\text{conf}}$ and **rlS+rlM**$_{\text{conf}}$ to refer to the case of using the concatenation of the handshape and the hand movement subunit probability posteriors depending on the different setups presented above.

The remainder of the experimental setup is the same as the main study.

Table 7.11 and Table 7.12 present the lexeme assessment and form assessment results. As expected, confidence score-based assessment does not really affect the performance.

Table 7.11 – F_1 score of the correct lexeme assessment using the posterior-based confidence measure Conf$_\alpha$ and using the single view or the multiple views reference models

	Production Space									
Reference Model	**rlS**$_{\text{conf}}$	**rlS**	**M**$_{\text{conf}}$	**M**	**rlM**$_{\text{conf}}$	**rlM**	**rlS+M**$_{\text{conf}}$	**rlS+M**	**rlS+rlM**$_{\text{conf}}$	**rlS+rlM**
Single view	0.73	0.73	0.84	0.84	0.79	0.79	0.83	0.83	0.78	0.80
Multiple views	0.71	0.72	0.88	0.88	0.85	0.85	0.88	0.90	0.87	0.87

Table 7.12 – F_1 score of the form error assessment using the posterior-based confidence measure Conf$_\alpha$ and using the single view or the multiple views reference models

		Production Space									
Reference Model	**Form**	**rlS**$_{\text{conf}}$	**rlS**	**M**$_{\text{conf}}$	**M**	**rlM**$_{\text{conf}}$	**rlM**	**rlS+M**$_{\text{conf}}$	**rlS+M**	**rlS+rlM**$_{\text{conf}}$	**rlS+rlM**
Single view	hshp	0.74	0.74	-	-	-	-	0.76	0.76	0.75	0.76
	hmvt	-	-	0.87	0.87	0.83	0.83	0.85	0.85	0.83	0.83
Multiple views	hshp	0.77	0.77	-	-	-	-	0.83	0.83	0.82	0.82
	hmvt	-	-	0.91	0.90	0.86	0.86	0.90	0.90	0.87	0.87

7.8 Demonstrator

The goal of the SMILE project was to develop an advanced platform which allows to assess Swiss German Sign Language (DSGS). Specifically, a sign language system that can assist Swiss

German Sign Language learners as well as aid in standardizing a vocabulary production test that can be aligned with levels A1 and A2 of the Common European Framework of Reference for Languages (CEFR). In this context, in collaboration with our SMILE project partners, a demonstrator that integrates the proposed sign language assessment approach and provides feedback to users was developed.

The system flowchart depicted in Figure 7.6 summarizes this system. Briefly a capture software extracts the 3D poses skeleton of the user sign production by using a Kinect camera. Then, these 3D poses with the video are fed to the handshape and the hand movement subunit probability estimators. These posterior features are time aligned by DTW algorithm with the KL-HMM-based references to obtain the assessment scores based on the SKL-divergence. Finally, a visualisation of the feedback based on these assessment results is provided to the user (details of the visualisation are presented later in the section).

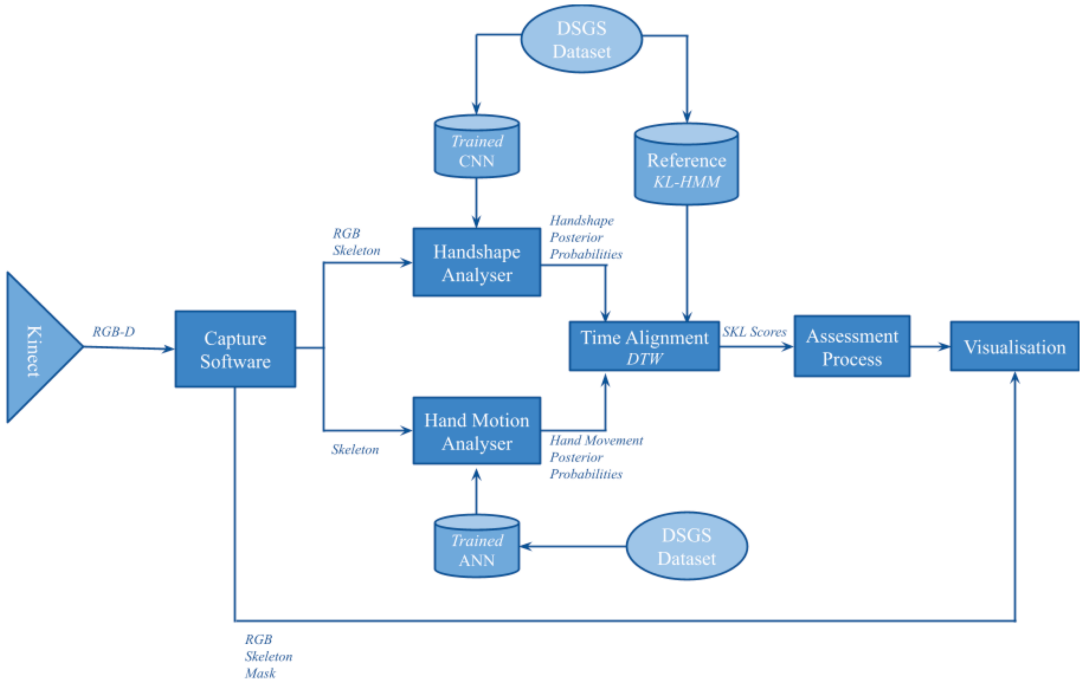

Figure 7.6 – Flowchart of the demonstrator of the project SMILE.

All the project partners contributed to the development of the demonstrator. More precisely,

- *DSGS Dataset*: creation of DSGS sign language resources and tools, and DSGS data collection were done by HfH and USurrey;
- *DSGS Dataset*: data annotation, lexicon management and the development of the assess-

ment famework was done by HfH;

- *Capture Software*: development of the data capture tools, body tracking and hand pose estimation was done by USurrey;
- *Handshape Analyser, Trained CNN*: development of the handshape estimator was done by USurrey;
- *Hand Motion Analyser, Trained ANN*: development of the hand movement estimator was done by us at Idiap;
- *Reference KL-HMM, Time Alignment, Assessment Process*: development of the sign assessment system was done by us at Idiap;
- *Visualisation*: development of the front-end software, feedback video and integration of all the components was done by USurrey.

HfH refers to the Hochschule für Heilpädagogik (in Zürich), USurrey refers to the University of Surrey (UK) and Idiap refers to the Idiap Research Institute (Martigny).

7.8.1 Assessment process of the demonstrator

The assessment process used in the demonstrator carries out assessment at form-level in time local manner. In other words, the demonstrator localizes the error to a specific segment of the sign such as a wrong handshape. To achieve that, the form assessment is carried out at KL-HMM state level as opposed to whole sign level, i.e.,

$$\mathscr{S}^{\text{multi}}_{form,f}(n) = \frac{\sum_{t=t_n^b}^{t_n^e} \text{SKL}(\mathbf{y}_{n,f}, \mathbf{z}_{t,f})}{t_n^e - t_n^b + 1} \,. \tag{7.11}$$

A threshold is then applied on each of the state scores, i.e. on $\mathscr{S}^{f}_{form}(n),\ \forall n$. This yields intervals of time frame which contain form error(s).

The demonstrator assesses four aspects of the isolated sign production: the handshape, the hand movement, the hand position and the hand location. The hand position refers to the position of the dominant and the non-dominant hand relative to each other, i.e. if the combination of the two hands is correct. This aspect is relevant for two-handed signs. For example, if for a right-handed signer, the left hand is upper than the right one in the sign NOM (name) in Swiss French Sign Language (see Figure 7.7), it is a hand position error since the left hand has to be below the right hand. The hand location refers to the localisation of the production of the sign in the signing space. For example, there is a hand localisation error if the sign BRAVO (well done) in Swiss French Sign Language is done at the chest-level instead of the head-level (see Figure 7.8).

Figure 7.7 – The sign production of the sign NOM (name) in Swiss French Sign Language

Figure 7.8 – The sign production of the sign BRAVO (well done) in Swiss French Sign Language

The proposed assessment approach detects the hand movement related error. But, it does not reveal if it is a hand movement, a hand position or a hand location error. To assess if it is a hand movement, a hand position or a hand location error, we integrated an additional criterion based on the (x,y) coordinate, called space-based criterion. More precisely, for each hand we compare the x-position of the user hand in the head coordinate center to the x-position of the reference sign (a cat.1 sample). As well as the y-position in the shoulder coordinate center (see Figure 7.9). If one x or y location differs from the reference sign, we label it as incorrect.

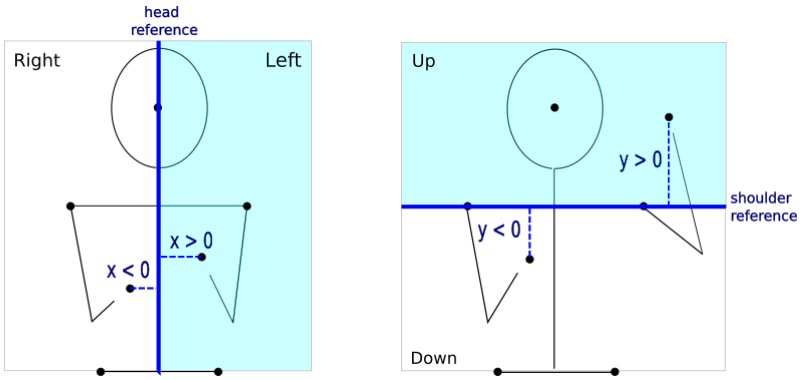

Figure 7.9 – Illustration of the additional assessment space-based criterion that allows to distinguish hand movement, hand position and hand location error

More precisely, there are two assessment criteria, one is time-based and the second is space-based. As shown in Figure 7.10: firstly the assessment scores are computed for the hand movement channel for the dominant and the non-dominant hand separately (written R for right-dominant and L for left-non-dominant in Figure 7.10). Then, the time frame interval $[a, b]$ in which the hand movement error appears is obtained, as explained above. When an error is detected in a frame interval $[a, b]$, the space-based criterion is applied to detect if the space location of each hand is correct or not. Given both criteria results, the final assessment is made in the following manner:

- **time-based**: R and L no error detected ⇒ no hand movement, hand position and hand location error;
- **time-based**: R (or L) error detected,
  - **space-based**: R and L no error detected ⇒ hand movement error;
  - **space-based**: R (or L) error detected ⇒ R (or L) hand position error;
  - **space-based**: R and L error detected ⇒ hand movement error or R (or L) hand position error;
- **time-based**: R and L error detected,
  - **space-based**: R and L no error detected ⇒ hand movement error;
  - **space-based**: R (or L) error detected ⇒ R or L hand position error;
  - **space-based**: R and L error detected ⇒ hand location error.

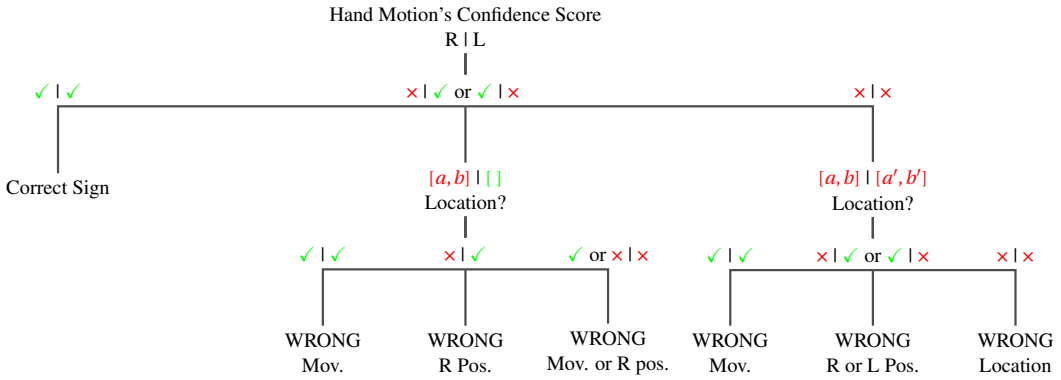

Figure 7.10 – Diagram of the integration of both, time-based and space-based, assessment criteria to determine if there is a hand movement, a hand position or a hand location error.

A production score of the handshape and the hand movement is computed based on the school grading mechanism, where in Switzerland the success threshold is $\frac{4}{6} \cong 0.66$. To obtain this score,

the total number of frames labelled as correct during the time-based criteria is counted and divided by the total number of frames for each channel, i.e. the handshape and the hand movement of each hand. A global production score is obtained based on these production scores by averaging them.

7.8.2 Front-end overview

The demonstrator front-end overview is depicted in Figure 7.11. After the welcome screen which is composed of a welcome video in DSGS, there are two modes proposed to the user: a practice mode (the left column of Figure 7.11) and a test mode (the right column).

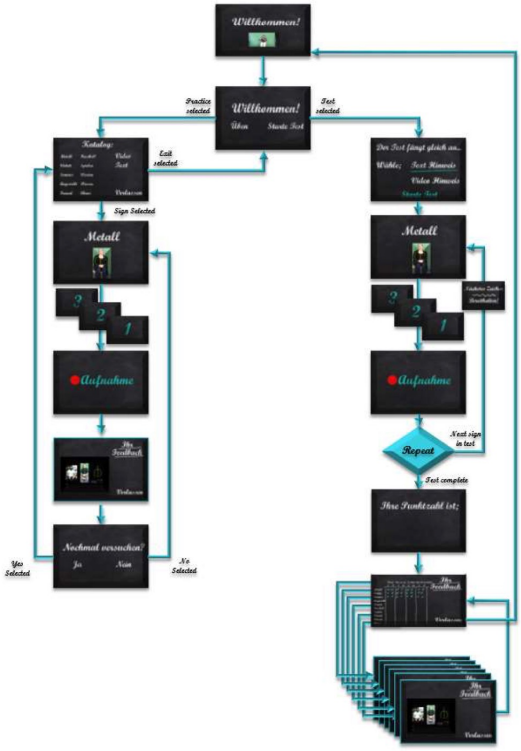

Figure 7.11 – Front-end overview of the demonstrator composed of a practice mode (left) and a test mode (right).

- In the practice mode, the user is directed to the lexicon catalogue, where the user can chose which sign she/he wants to practice. Three signs were available, namely VIOLETT (violet), SOMMER (summer) and METALL (metal). The user can also choose at this stage if she/he wants to see an example of the reference video before practicing. The next step plays the reference video if the option was chosen. Then a countdown finishes on the video recording

of the user with the capture tool. After processing a detailed feedback, see Figure 7.12, is provided to the user. The feedback screen includes

- on the left: the video of the user sign production with the joint skeleton drawn on it;
- in the middle: a reference video with also the joint skeleton drawn on it;
- on the right: both skeletons aligned with below the production scores of both handshape and both hand movement;
- down: a time line of the production of the reference and the user;

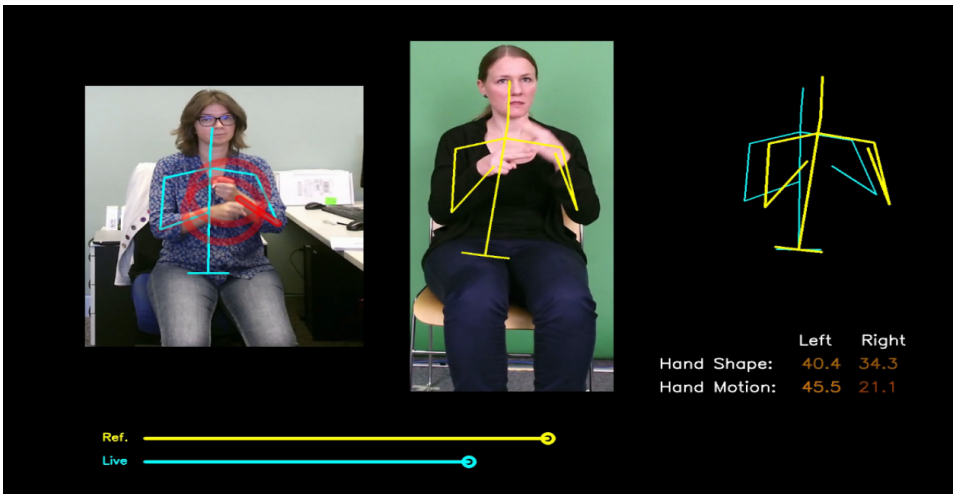

Figure 7.12 – Detailed feedback screen of the demonstrator.

A per-frame feedback is projected on the user production video with a red circle around the hand when the handshape is incorrect and a red line on the forearm when the hand movement is incorrect. The circles/lines are coloured based on the SKL scores. A per-hand score is given by the production scores and a feedback on the speed production is given by the time lines of the production.

- In the test mode, the users are asked to record each sign of the catalogue which are randomly selected. Before starting it, the user has the possibility to choose the video option to see an example of the reference video, before each recording. At the end of the test, a feedback screen, see Figure 7.13, is to the user which is a feedback table on the handshape (Hand), the hand location (Bewegung), the hand position (Position) and the hand movement (Geschwindigkeit) of each hand where each sign is an entry. A tick or a cross gives the assessment for each column. Besides the table, the average of the global production score of each sign gives the success percentage of the test. Furthermore, the user can select each sign to see the detailed feedback screen presented in the practice mode (see Figure 7.12).

Figure 7.13 – Overview feedback screen of the test mode of the demonstrator.

The demonstrator was tested on the available linguistically annotated data and was presented to the public[1].

7.9 Summary

This chapter presented a phonologically motivated sign language assessment approach that allows to assess two different linguistic aspects of a produced sign: the lexeme and the form. In this approach, a produced sign is matched to a reference model and a decision is made based on the best matching path. Two reference models were developed: a multiple views reference model based on KL-HMM and a single view reference model based on a single instance (one shot) of an acceptable production of the sign. A validation study on the SMILE DSGS database yielded promising lexeme-level assessment and form-level assessment results. We found that, although the multiple views reference model yields better performance, the single view reference model gives relatively good performance, despite the fact that the reference is based on a single instance of sign production. Our studies also showed that the different components of the proposed assessment system can be built only using cat.1 and cat.2 data. Further analysis studies investigating with clustered hand movement subunits and selection of HMM topology through model selection showed that those developments extend to sign language assessment task. The proposed sign language assessment was successfully integrated into a real-time demonstrator that assesses isolated sign production at form-level and provides feedback.

[1] https://www.idiap.ch/project/smile/news/smile-how-it-works (visited on 02/19/2021)

8 Conclusion and Future Directions

The goal of the thesis was to develop an explainable framework for sign language recognition and assessment, where the sign language is acquired through a camera. To do so, we focused on developing a HMM-based framework that can carry out recognition/verification and linguistically valid assessment in an integrated manner. One of the main motivations behind using HMM was that HMM allow integration of both prior knowledge and data, and makes the system modular and interpretable. Having said that, when applying HMM some level of prior knowledge is needed to determine the topology. Such prior knowledge is not readily available for sign languages. So, as a first step, in Chapter 4, this thesis focused on addressing this challenge by proposing a model selection approach, where each sign is modeled by HMM with different number of states and the system infers the most likely topology during the recognition phase. This approach was found to yield better system when compared to presetting the HMM topology based on k-means. In the later part of the thesis, it was found that a similar model selection criterion can be applied on the development data to determine the HMM topology.

In recent years, exploiting the discrete nature of handshape, different ways to model handshape information for sign language recognition has emerged. However, a sign is not entirely defined by handshape. There are other manual channels such as, hand movement that needs to be modeled. One of the main challenges in modeling hand movement channel is that it is continuous in nature. We addressed that challenge in Chapter 5 to develop methods to model hand movement as discrete subunits. More precisely, we developed an approach where, given only pairwise comparison between sign productions, hand movement subunits are derived from 3D skeleton information by building sign-level HMM and clustering those HMM states through a measure of discrimination. Our studies showed that both the sign-level HMM and clustered set of HMM states can serve as discrete hand movement subunits. Furthermore, cross-lingual and multilingual sign language recognition studies showed that these subunits are transferable across sign languages. Thus, paving the path to model hand movement information exploiting multiple sign language resources, like modeling of handshape information. In addition, we also showed that the proposed

approach for subunits derivation is abstract. In the sense that it can be applied to other problems, like in speech processing to discover subword units or phone set and develop pronunciation lexicon. As part of the investigations to understand the derived hand movement subunits, we also developed a visualization approach inspired from movement synthesis in 3D space in robotics.

Sign language consists of different channels of information corresponding to manual components (hand position, hand movement and handshape) and non-manual components (mouthing, facial gesture, posture). Besides extraction of the information related to these components from the visual signal, there is need for methods that can effectively model these components jointly for sign language processing. In that respect, we argued that this challenge is similar to the challenge of modeling multichannel speech production (phonological) information or articulatory features in speech processing. In Chapter 6, we showed that the HMM-based methods developed in speech processing for modeling articulatory features can be adopted for modeling the multichannel information in sign languages. We proposed two phonology-based approaches to model jointly different visual subunits: tandem-feature based approach and KL-HMM based approach. Through extensive studies on modeling hand movement and handshape channels for monolingual, cross-lingual and multilingual sign language recognition, we showed that joint modeling of hand movement and handshape channels through both the approaches consistently improves over stand-alone hand movement channel modeling and stand-alone handshape channel modeling.

Chapter 7 built on the phonological framework developed in Chapter 6 for joint modeling of multichannel information to develop an explainable sign language assessment approach. Specifically, we showed that the KL-HMM based phonological approach can be naturally extended for sign language assessment. This approach carries out assessment of sign at two different levels: at lexeme level and at form level. Extensive studies on SMILE DSGS database showed that (a) with the proposed approach sign language assessment can be carried in a linguistically valid manner, (b) the KL-HMM can be replaced by a single instance of acceptable sign production of the expected sign, and (c) interpretable assessment scores can be generated. The investigations in this chapter also led to development of a real-time demonstrator where isolated sign productions are assessed at form level and feedback is provided. It is worth mentioning that it is one of the first works where hand movement and handshape are automatically assessed at different levels.

There are potential directions for future research,

- This thesis focused on modeling two channels in sign language: hand movement and handshape. However, phonology-based sign language recognition and assessment approaches developed in this thesis as such are not restrictive to handshape and hand movement information. Other information such as facial expression, mouthing could be modeled by feature augmentation. One such potential direction could be the facial action coding system [25] which describes the facial muscular positioning into basic universal emotions. Such system

can give relevant information on eye/eye-brown movements, mouth movements, head movements and emotions/expressions.

- The research and development in this thesis focused only on isolated signs. However, in sign language communication, similar to spoken language, there is use of continuous signing. When compared to isolated sign production, one of the differences in continuous signing is that there is a "co-articulation" effect when transiting from one sign/lexeme to another sign/lexeme. Thus, further research is needed to scale the sign language recognition and assessment approaches developed in this thesis to continuous signing. One potential way to handle the co-articulation effect between the signs is to use the transition model proposed in Chapter 4.

- In this thesis, we have mainly used an acquisition system that is based on Kinect camera. However, in the recent years, in the computer vision community RGB cameras have gained increased attention and considerable progress has been made in the area of human motion/action recognition [121]. Scaling of the proposed approaches to a RGB camera acquisition system is open for future research.

- As demonstrated in Chapter 5, hand movements can be synthesized in the 3D skeleton space by using the subunits-based HMM inferred for signs as a generative model. The hand movement synthesis could potentially be used as input to a sign production system such as an avatar system [37] or a neural-based sign production system [107, 106]. Another area of interest could be robotics, where a robot performing manual component of sign could be conceived through generation of both hand movement and handshape information.

- The sign language assessment system developed in this thesis presumes that the signs are produced by adult signers. Computer-aided sign language learning tool can be of interest to children as well. So, an interesting research question arises is how to scale such a system to children?

A Subunits Extraction for Spoken Language Application

State-of-the-art methods for development of Automatic Speech Recognition (ASR) systems and Text-to-speech Synthesis (TTS) systems presume that the target language has a written form and there exists a phonetic lexicon that transcribes the written form into sequence of phonemes/phones. Given the written form of words, a phonetic lexicon can be developed with the help of linguistic expertise or knowledge of the target language [1, 56]. As a first step, a human expert manually transcribes each word into a phoneme sequence by observing the grapheme sequence (i.e. orthographic transcription). Once a base lexicon is available, a rule-based approach [30, 31] or a learning-based approach (e.g., grapheme-to-phoneme conversion [82, 11, 119]) can be adopted to augment the lexicon with new words and pronunciation variants.

In the world, there are approximately 6900 languages and only about 5-10% of them employ a writing system [1]. Furthermore, not all of the languages that have a writing system may have a well developed phonetic dictionary. Studying these languages manually, to acquire linguistic knowledge and phonetic dictionary from the acoustic data, is a highly challenging and non-trivial task. Availability of computational methods can immensely help both the linguistic research community as well as the speech technology research community. One potential venue for that is the area of zero-resource speech processing,[1] which originally started with the problem of unsupervised speech pattern discovery [83, 84], and was then extended to automatic subword units discovery [113] and spoken term discovery [50], and more recently cast as a problem of automatic language acquisition by machines [114, 32], taking inspirations from how infants and children acquire spoken language at the very early stages of life.

Instead of a completely unsupervised approach, yet another approach could be addressing a somewhat simplified question with light supervision: *given only a set of utterances and the knowledge that any two pair of utterances correspond to the same word or not, how to*

[1] https://zerospeech.com/2015/index.html (visited on 02/19/2021)

automatically infer the phone set inventory and a lexicon? Irrespective of whether the target language is known or not or whether it has a written form or not this question can be posed. Furthermore, linguish notions such as minimal pairs are built on pairwise comparison. We can pose subword unit extraction in a similar manner. Availability of such a data with pairwise comparisons can very well be envisaged in field linguistics. For instance, collection of speech utterances of day-today life objects/entities (e.g. food, cloth, numbers) possibly without the necessity to speak the unknown spoken language by showing them. Also, if only acoustic data of an unknown language is available, such form of light supervision, i.e. whether two utterances correspond to the same word or not, could be obtainable from people with some speech expertise through listening tests. Furthermore, such a question can be posed in the above mentioned zero resource speech processing framework after unsupervised spoken term for phone set or automatic subword unit inventory discovery and pronunciation model extraction. The same question could be posed in the case of sign languages to derive subunits and model signs as a sequence of subunits as presented in Chapter 5.

In this appendix, we applied the HMM-based abstract framework presented in Chapter 5 for linguistic resource development for speech processing, by building upon the inherent ability of HMM to segment time series into stationary segments and recent works on resource-constrained speech processing using auxiliary multilingual resources. We demonstrate with the spoken language study that the framework can lead up to phone set discovery and pronunciation lexicon development.

A.1 Spoken Subunit derivation and Lexicon Development

This section presents the subunit based lexicon development (see Chapter 5) applied to spoken languages. Specifically, given the pairwise comparison data, in this methodology,

Step 1: first, a sequence of feature vectors is extracted for each utterance. The feature vectors are short-term cepstral features, which tend to model information related to vocal tract system;

Step 2: given the sequence of feature vectors for each utterance, a HMM is obtained for each unknown word in the set;

Step 3: next, the states are clustered into subunits by pairwise comparison and a sequence model in terms of clustered subunits is obtained for each unknown word; and finally

Step 4: phone set and pronunciation model for the unknown words are inferred by learning a probabilistic subunit-to-phone relationship exploiting auxiliary speech data with linguistic resources.

Methodology of steps 2 and 3 remain the same as in the sign language application; i.e. in Step 2 word-level HMM are determined and given the single Gaussians of the words HMM states, Step 3 clusters them through the Bhattacharrya distance, a measure of discrimination.

In Step 4, the goal is to establish a link to linguistic knowledge to ascertain the identity of the automatic subword units. Since spoken language can be written in terms of phones, this can be done by learning a probabilistic relationship between the derived subword units and phones through acoustic signal. More precisely, as illustrated in Figure A.1, this involves,

1. training of a phone posterior probability estimator on auxiliary data or languages that have well-developed phonetic resources;
2. training of KL-HMM [4, 3] with the phone posterior probability as feature observations and the states being represented by the derived automatic subword units. Each state of the KL-HMM is parameterized by a categorical distribution of the same dimension as phone probability feature vector, which capture a probabilistic relationship between the automatic subword units and the phones; and
3. inference of phone-based pronunciation by using the trained KL-HMM as a generative model and decoding the resulting sequence of probability through an ergodic HMM of phones.

It is worth mentioning that the proposed approach of inferring phonetic identities of the automatic subword units, and consequently a phonetic lexicon is inspired from the approach of acoustic data-driven grapheme-to-phoneme conversion using KL-HMM [95].

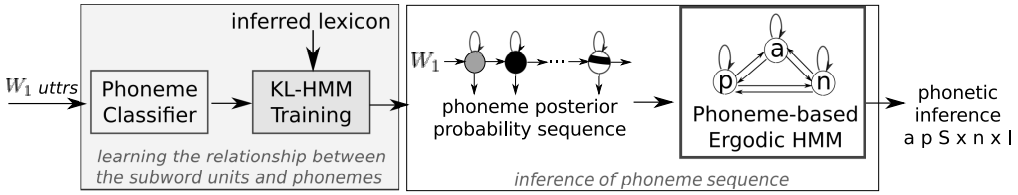

Figure A.1 – Illustration of the phoneme inference according to the derived subword units.

A.2 Experimental Setup

We validated the proposed approach on a spoken language through ASR system level studies, as ASR relies on discrimination between words, and pronunciation lexicon level studies. We used a part of the PhoneBook database for the study. We used 39 dimensional Perceptual Linear Prediction (PLP) [46] cepstral features ($c_0 - c_{12} + \Delta + \Delta\Delta$) extracted with a window size of 25 ms and with a window shift of 10ms as the feature vectors.

A.2.1 PhoneBook database

PhoneBook is a speaker-independent phonetically-rich isolated-word telephone-speech English database [89]. PhoneBook consists of more than 92,000 utterances and almost 8,000 different words, with an average of 11 different speakers/word. The database has been split into 106 word lists, each composed of around 75 words. Furthermore, the set of speakers is different for each word list. The word list contains uncommon English words and proper names (e.g., Witherington, Gargantuan). For our investigation, we used the small size (75 words) vocabulary setup; more precisely the *ad* word list that we separated into training, development and test sets as follows:

We selected speaker *m0k* who has uttered 74 words out of the 75 words as the development set. The development set is used for determining the number of HMM states per word in Step 2 and the clustering threshold τ in Step 3. With the ten remainder speakers, we performed a speaker independent experiment, where a leave-one-speaker out protocol was applied. Thus, ten experiments were conducted, where in each experiment, the data of one speaker was used for testing and the data of the remaining speakers are used for automatic subword units inference and for training ASR system. For each of the experiment, the average number of utterances for training, development and testing are 621, 74 and 69, respectively.

For lexical level validation studies, as part of Step 4, we used 21 word lists: *aa, ah, am, aq, at, ba, bh, bm, bq, bt, ca, ch, cm, cq, ct, da, dh, dm, dq, dt, ea* to train phone-based classifier. This word list was originally defined in a study on task-independent speaker-independent speech recognition [33]. Task-independent because the words in each word list are different and speaker-independent as the speakers in each word list are different. For example, words and speakers in word list *ad* are entirely different than the words and speakers in the 21 word lists. As done in [33], we use 42 context-independent phones (including silence) from the PhoneBook dictionary.

We also conducted a study where the phone posterior probability estimator is trained with multilingual data without English. For that we used the Swiss French, Swiss German, Italian and Spanish part of the SpeechDat(II) corpus. Each language has about 12 hours of speech. The lexicons are based on SAMPA phone set.[2] We created a multilingual phone set by merging the phone sets across the four languages. This resulted in 104 context-independent phones including silence. It is worth mentioning that 35 phones out of the 104 phones are common to English SAMPA phone set.

A.2.2 Systems

We built HMM/GMM [92] and hybrid HMM/ANN [14] systems to evaluate the automatic subword units based lexicon at ASR level. We built KL-HMM systems for lexical level validation.

[2] https://www.phon.ucl.ac.uk/home/sampa/ (visited on 02/19/2021)

In each case, we built two systems: (a) using word-level HMM states obtained in Step 2 as subword units, referred to as word level system and (b) using the clustered HMM states in Step 3 as subword units, referred to as clustered subword units based system. The motivation behind building word level system is that Step 2 obtains a word level HMM with fixed number of states n through discrimination like in Step 3, so the states of the word level HMM can also regarded as subword units without being clustered. Such a comparison would help us to determine whether the clustering step is indeed yielding meaningful subword units or not. The HMM were trained and tested with the HTK toolkit [130]. The KL-HMM system studies were conducted using an in-house modified version of HTK. The neural networks, more precisely MLP, for hybrid HMM/ANN and KL-HMM systems were trained using the Quicknet software [52].

HMM/GMM Systems: All the HMM/GMM systems are left-to-right HMM using a single Gaussian distribution with diagonal covariance matrix as the emission distribution. In the case of the word level system, the number of states is chosen according to the model that saturates on the training and development data (Step 2). The range of the number of states was from 3 to 30. In the subword unit-based model, the clustering step was conducted with the hyper-parameter, τ, in the range of 0.8 to 3.2 with a 0.2 step, each leading to a different lexicon. An HMM/GMM system was trained for each lexicon and the final one was chosen according to the maximum recognition accuracy on the development set (Step 3). The resulting word level system and clustered subword unit based system was tested on the test set. This process was repeated for each speaker-independent fold.

Hybrid HMM/ANN Systems: For building the hybrid HMM/ANN systems, we first obtained the alignments in terms of the HMM states respectively from the word level and the clustered subword units-based HMM/GMM systems for each speaker-independent fold. We then trained MLPs classifying HMM states with output non-linearity of softmax and minimum cross-entropy error criterion. We used the 39-dimensional PLP cepstral features with four frames preceding context and four frames following context as the MLP input. In our experiments we trained MLPs with different number of hidden units (600, 800, 1000) and hidden layers (0, 1, 2, 3). The number of hidden units and hidden layers as well as other hyper-parameters such as learning rate and the batch size were chosen according to the frame-level accuracy on the development set.

We estimated the scaled likelihoods in the hybrid HMM/ANN systems by dividing the posterior probabilities derived from MLPs with the prior probabilities of the classes estimated from relative frequencies in the training data. These scaled likelihoods were then used as emission probabilities for HMM states.

KL-HMM Systems: First a single hidden layer MLP was trained to classify 42 context-independent phones, including silence. We used the 39-dimensional PLP cepstral features with four frames preceding context and four frames following context as the MLP input. The number of hidden nodes was 800. The KL-HMM parameters were then training by forward passing the training portion of the *ad* list data through the MLP and using resulting 42 dimensional phone posterior probability distribution per frame as the feature observations. We trained word level system and clustered subunits based system for each speaker-independent fold. After training the KL-HMM system, we tested the performance at ASR level on the test data. For lexical level validation, we generated the pronunciation of each word in terms of the 42 phones, as described earlier in Step 4. For each word, we then computed the Levenshtein distance between the inferred pronunciation and the pronunciation given in the PhoneBook dictionary.

We trained a multilingual KL-HMM system for each speaker-independent fold where, we first trained a multilingual phone classifier on the SpeechDat(II) and then for each fold trained the KL-HMM parameters on the training portion of the *ad* list data, by forward passing it through the multilingual phone classifier and using the resulting multilingual phone posterior probabilities as feature observation.

A.3 Results and Analysis

In this section, we first present ASR level validation studies followed by lexical level validation studies. We then present as part of analysis: (i) impact of number of utterances on the proposed methodology on phone set and pronunciation model inference and (ii) investigations using language independent multilingual data.

A.3.1 Automatic speech recognition level validation

First, the ASR study on the PhoneBook database is conducted to validate the assumption that the proposed approach derived discriminative subword units. Table A.1 presents the average RA over the ten fold experiments of the clustered subword unit-based and word level systems as well as the average number of units used per system. It can be observed that, in the case of HMM/GMM

Table A.1 – Clustered subword unit-based and word level systems RA on the PhoneBook database using PLP cepstral features with HMM/GMM and hybrid HMM/ANN systems

	Clustered subword unit-based System	Word level System
HMM/GMM	94.1 ± 5.6	96.1 ± 4.0
Hybrid HMM/ANN	97.8 ± 2.0	98.3 ± 2.1
Average # units	*810*	*1125*

study, word-level system outperforms clustered subword units based system. However, in the case of hybrid HMM/ANN system, the performances are better than respective HMM/GMM system performance and are comparable. As a whole, the results indicate that the clustered subword units retain discrimination information across the words even with a reduction of around 28% of the number of HMM states.

Table A.2 presents the average RA of the clustered subword unit-based and word level KL-HMM systems. As it can be seen, both systems yield comparable RAs, again indicating that clustered subword units retain discrimination across the words.

Table A.2 – KL-HMM-based subword unit- and word level -based systems results on the PhoneBook database using posterior distributions as features

	Clustered subword unit-based system	Word level system
KL-HMM	99.0 ± 1.8	99.4 ± 1.2

A.3.2 Lexical level validation

As explained earlier in Section A.1 (see Figure A.1), we inferred the pronunciation of each word in the lexicon using the KL-HMM as a generative model, and decoding the resulting sequence of phone posterior probabilities for each word using a 42 phone fully connected ergodic HMM to get the pronunciation model. We compared the inferred pronunciations with the pronunciation provided in the PhoneBook dictionary, and computed Levenshtein distance (LEV) [67] and Phone Recognition Rate (PRR). PRR is calculated as

$$1.0 - \frac{(\#insertion + \#deletion + \#substitution)}{N_{ref}}, \tag{A.1}$$

where # denotes 'number of' and N_{ref} denotes the number of phones in the reference phonetic transcription. Table A.3 presents the average LEV and PRR for pronunciations inferred by clustered subword unit based KL-HMM and word level KL-HMM. It can be observed that the inferred pronunciations are close to the original pronunciations in the manual pronunciation dictionary. Further analysis of the lexicon showed that the phonetic lexicon inferred using clustered subword unit based system cover 39 phones out of the 42 phones, while the manual dictionary for the words in *ad* list covers 38 phones. More precisely, with an exception of one extra phone, all the phones in the manual dictionary were inferred.

Table A.3 – Levenshtein distance (LEV) and phone recognition rate (PRR) results of the lexicon inferred from clustered subword unit-based KL-HMM system and word level KL-HMM system

	Clustered subword unit-based system	Word level system
LEV ± std	1.9 ± 0.2	1.5 ± 0.1
PRR ± std	70.3 ± 2.6	76.4 ± 1.1

A.3.3 Further analysis

Impact of number of utterances: In the experiments presented above, we had nine speakers utterances per word to derive subword units. In realistic under-resourced language scenario, it may not be possible to get so many speaker utterances per word. So we studied the impact of number of speaker utterances on the proposed approach by developing two additional systems: (a) using only six speaker utterances per word (denoted as *six-utterances*) and (b) using only four speaker utterances per word (denoted as *four-utterances*) in a gender balanced manner. We compared the performances to the case where all the training utterances (denoted as *all-train-utterances*) are used. It is worth mentioning that after deriving automatic subword unit lexicon the HMM/GMM system was trained with all the utterances so that we can fairly compare the resulting lexicons. If the HMM/GMM systems were trained with four utterances or six utterances, separating the differences due to lexicon and data sparisty would have been a non-trivial task. Table A.4 presents the average RA for HMM/GMM system. It can be observed that the amount of data used to infer automatic subword unit based pronunciation lexicon does not seems to affect the performance at ASR level. Interestingly, we can observe improvement with *six-utterances* based lexicon.

Table A.4 – Clustered subword unit-based and word level HMM/GMM systems results on the PhoneBook database depending on the three different setups used to infer the lexicon (*all-train-/six-/four-utterances* based lexicon) using PLP cepstral features

	HMM/GMM-based system		
	Lexicon	Average RA ± std	*Average # units*
Clustered subword unit-based system	*all-train- utterances*	94.1 ± 5.6	*810 (-28%)*
	six-utterances	95.7 ± 4.5	*1365 (-9%)*
	four-utterances	95.4 ± 5.9	*1019 (-3%)*
Word level system	*all-train-utterances*	96.1 ± 4.0	*1125*
	six-utterances	96.3 ± 4.0	*1500*
	four-utterances	96.0 ± 5.5	*1050*

Table A.5 presents the results with KL-HMM system at ASR level and lexical level. In this case, the automatic subword units based lexicon is derived using *all-training-, four-* or *six- utterances* and the KL-HMM is also trained on *all-train-, four-* or *six- utterances*, respectively. We can again observe that reduction in number of utterances is not affecting Step 2, Step 3 and Step 4. Similar

to the HMM/GMM study, at the ASR level, the number of samples are not impacting the RA. At the LEV and PRR level, we can observe improvement with the *six-utterances* based lexicon.

Table A.5 – Clustered subword unit-based and word level KL-HMM systems RA, Levenshtein distance (LEV) and phone recognition rate (PRR) on the PhoneBook database depending on the three different setups used to infer the lexicon (*all-train-/six-/four-utterances* -based lexicon)

	Monolingual KL-HMM-based system			
	Lexicon	Average RA ± std	LEV ± std	PRR ± std
Clustered subword unit-based system	*all-train-utterances*	99.0 ± 1.8	1.9 ± 0.2	70.3 ± 2.6
	six-utterances	99.3 ± 1.2	1.5 ± 0.1	76.2 ± 1.7
	four-utterances	99.0 ± 1.4	1.8 ± 0.1	71.5 ± 1.0
Word level system	*all-train-utterances*	99.4 ± 1.2	1.5 ± 0.1	76.4 ± 1.1
	six-utterances	99.3 ± 1.2	1.4 ± 0.0	77.5 ± 0.7
	four-utterances	99.1 ± 1.4	1.8 ± 0.1	72.1 ± 0.8

Multilingual study: In the above studies, the ASR level and lexical level studies were conducted in matched condition in term of language. In other word although the words and the speakers are not shared across word lists, the language is still English. So we studied the possibility to use language-independent multilingual resources. For that, we performed ASR and pronunciation inference study using the multilingual KL-HMM system, where the 104 dimensional multilingual posterior probabilities estimated by MLP trained on Swiss French, Swiss German, Italian and Spanish portions of SpeechDat(II) are used as the feature observation to learn the relationship between the automatic subword units and the multilingual phones. Table A.6 presents the ASR performance in terms of RA. For all the case, there is slight drop in performance when compared to the monolingual MLP. The trend remains same word level system and clustered subword units based system yield comparable systems. We also observe that reducing the number of utterances for derivation of automatic subword units and KL-HMM training does not impact the performance of the systems. This suggests that there exists a systematic relationship between the derived subword units and multilingual phones.

Table A.6 – Clustered subword unit-based and word level KL-HMM systems results on the PhoneBook database using multilingual phoneme classifier (without English)

	Multilingual KL-HMM-based system	
	Lexicon	Average RA ± std
Clustered subword unit-based system	*all-train-utterances*	98.4 ± 1.5
	six-utterances	98.5 ± 1.6
	four-utterances	98.4 ± 2.5
Word level system	*all-train-utterances*	98.7 ± 2.1
	six-utterances	98.5 ± 1.6
	four-utterances	98.4 ± 2.5

We inferred the pronunciation based on the 104 multilingual SAMPA phone set. We found that *all-train-utterances* based lexicon covers 40 phones out of the 104 phones. 27 out of the inferred 40 phones belong to or shared to English SAMPA phone set, while 13 are borrowed from other languages. We were not able to carry out lexical level validation using LEV and PRR measures, as PhoneBook lexicon and SpeechDat(II) lexicon are based on two different Bets. It was not possible to map all the phones precisely, especially multilingual phones. Table A.7 presents some examples of the phonetic inference in PhoneBook Bet for the monolingual case and SAMPA Bet for the multilingual case. We can observe that, unlike monolingual inference, the multilingual phone inference is somewhat noisy. This can potentially be due to the mismatch in the database conditions.

Table A.7 – Examples of phonetics inference according to the monolingual KL-HMM and multilingual KL-HMM

Word	True	Monolingual-based Inferrence	Multilingual-based Inferrence
yarns	y a r n z	y a r n z	j o n
speechwriter	s p i C r Y t X	s p i C r Y t X	s p i tS u a OY e l
infrequently	I n f r i k w x n t l i	I n f r i k w x t l i	i e n f w i k u e
oops	u p s	w u p t s	n u
quail	k w e l	k w e l	u w e i o
bonbon	b a n b a n	b @ a n b a x n	o n b o n

Bibliography

[1] M. Adda-Decker and L. Lamel. “Multilingual dictionaries”. In: *Multilingual Speech Proceesing*. Ed. by Tanja Schultz and Katrin Kirchhoff. Academic Press, 2006. Chap. 5, pp. 123–168.

[2] F. Amauger et al. *Langue des signes Francaise - A1*. Belin, 2013. ISBN: 9782701165677.

[3] G. Aradilla, H. Bourlard, and M. Magimai.-Doss. “Using KL-based acoustic models in a large vocabulary recognition task”. In: *Proc. of Interspeech*. 2008.

[4] G. Aradilla, J. Vepa, and H. Bourlard. “An acoustic model based on Kullback-Leibler divergence for posterior features”. In: *Proc. of the IEEE Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 2007.

[5] O. Aran. “Vision based sign language recognition: modeling and recognizing isolated signs with manual and non-manual components”. PhD thesis. Istanbul, Turkey: Bogazici University, 2008.

[6] M. Assan and K. Grobel. “Video-based sign language recognition using Hidden Markov Models”. In: *Gesture and Sign Language in Human-Computer Interaction*. Ed. by Ipke Wachsmuth and Martin Fröhlich. Springer Berlin Heidelberg, 1998, pp. 97–109. ISBN: 978-3-540-69782-4.

[7] G. Awad, J. Han, and A. Sutherland. “Novel boosting framework for subunit-based sign language recognition”. In: *Proc. of the 16th IEEE International Conference on the Image Processing (ICIP)*. 2009.

[8] B. Bauer and K.-F. Kraiss. “Towards an automatic sign language recognition system using subunits”. In: *Gesture and Sign Language in Human-Computer Interaction: International Gesture Workshop*. 2002. ISBN: 978-3-540-47873-7.

[9] A. Bhattacharyya. “On a measure of divergence between two multinomial populations”. In: *Sankhyā: The Indian Journal of Statistics (1933-1960)* 7.4 (1946), pp. 401–406. ISSN: 00364452.

[10] S. Bilal et al. “Dynamic approach for real-time skin detection”. In: *J. Real-Time Image Process.* 10.2 (2015), pp. 371–385. ISSN: 1861-8200. DOI: 10.1007/s11554-012-0305-2. URL: https://doi.org/10.1007/s11554-012-0305-2.

[11] M. Bisani and H. Ney. “Joint-sequence models for grapheme-to-phoneme conversion”. In: *Speech Communication* 50.5 (2008), pp. 434–451.

[12] R. E. Blahut. “Hypothesis testing and information theory”. In: *IEEE Trans. on Information Theory* IT-20.4 (1974).

[13] M. Bohner and N. Wintz. “The linear quadratic tracker on time scales”. In: *International Journal of Dynamical Systems and Differential Equations (IJDSDE)* 3.4 (2011).

[14] H. Bourlard and N. Morgan. *Connectionist speech recognition: a hybrid approach*. Norwell, MA, USA: Kluwer Academic Publishers, 1993. ISBN: 0792393961.

[15] R. Bowden et al. “A linguistic feature vector for the visual interpretation of sign language”. In: *Proc. of the European Conference on Computer Vision (ECCV) 2004*. Springer, pp. 390–401.

[16] D. Brentari, J. Fenlon, and K. Cormier. *Sign language phonology*. July 2018. DOI: 10.1093/acrefore/9780199384655.013.117. URL: https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-117 (visited on 02/17/2021).

[17] N. C. Camgöz, A. A. Kındıroğlu, and L. Akarun. “Gesture recognition using template based random forest classifiers.” In: *ECCV Workshops (1)*. 2014.

[18] N. C. Camgöz, A. A. Kındıroğlu, and L. Akarun. “Sign language recognition for assisting the deaf in hospitals”. In: *Proc. of the Human Behavior Understanding: 7th International Workshop*. 2016. ISBN: 978-3-319-46843-3.

[19] N. C. Camgöz et al. “BosphorusSign: A Turkish Sign Language recognition corpus in health and finance domains”. In: *Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC), Portorož, Slovenia, May 23-28, 2016*. 2016. URL: http://www.lrec-conf.org/proceedings/lrec2016/summaries/70.html.

[20] N. C. Camgöz et al. “Subunets: End-to-end hand shape and continuous sign language recognition”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2017.

[21] Z. Cao et al. “Realtime multi-person 2D pose estimation using part affinity fields”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[22] Ö. Çetin et al. “Monolingual and crosslingual comparison of tandem features derived from articulatory and phone MLPs”. In: *Proc. of Automatic Speech Recognition and Understanding Workshop*. 2007.

[23] J. Christopher. "SignAssess – Online sign language training assignments via the browser, desktop and mobile". In: *Computers Helping People with Special Needs*. Ed. by Klaus Miesenberger et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 253–260. ISBN: 978-3-642-31534-3.

[24] J. Coerts. "Nonmanual grammatical markers: An analysis of interrogatives, negations and topicalizations in Sign Language of the Netherlands". PhD thesis. University of Amsterdam, 1992.

[25] J. Cohn, Z. Ambadar, and P. Ekman. "Observer-based measurement of facial expression with the Facial Action Coding System." In: *The Handbook of Emotion Elicitation and Assessment*. 2007, pp. 203–221.

[26] T. F. Connor et al. "The language of glove: Wireless gesture decoder with low-power and stretchable hybrid electronics". In: *PLOS ONE* 12.7 (July 2017), pp. 1–12. DOI: 10.1371/journal.pone.0179766. URL: https://doi.org/10.1371/journal.pone.0179766.

[27] H. Cooper and R. Bowden. "Large lexicon detection of sign language". In: *Human–Computer Interaction*. Ed. by Michael Lew et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 88–97. ISBN: 978-3-540-75773-3.

[28] H. Cooper, B. Holt, and R. Bowden. "Sign language recognition". In: *Visual Analysis of Humans, 2011*. 2011. ISBN: 978-0-85729-997-0. DOI: 10.1007/978-0-85729-997-0_27. URL: http://dx.doi.org/10.1007/978-0-85729-997-0_27.

[29] H. Cooper et al. "Sign language recognition using sub-units". In: *Journal of Machine Learning Research 13* (2012).

[30] M. Davel and E. Barnard. "Pronunciation prediction with Default&Refine". In: *Computer, Speech & Language* 22 (2008), pp. 374–393.

[31] M. Dedina and H. Nusbaum. "PRONOUNCE: A program for pronunciation by analogy". In: *Computer, Speech & Language* 5 (1991), pp. 55–64.

[32] E. Dunbar et al. "The zero resource speech challenge 2017". In: *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2017, pp. 323–330.

[33] S. Dupont et al. "Hybrid HMM/ANN systems for training independent tasks: Experiments on 'Phonebook' and related improvements". In: *Proc. of ICASSP*. 1997.

[34] S. Ebling et al. "SMILE Swiss German Sign Language dataset". In: *Proc. of the LREC*. 2018.

[35] R. Elakkiya. "Machine learning based sign language recognition: A review and its research frontier". In: *Journal of Ambient Intelligence and Humanized Computing* (2020). DOI: 10.1007/s12652-020-02396-y. URL: https://doi.org/10.1007/s12652-020-02396-y.

[36] R. Elakkiya and K. Selvamani. “Extricating manual and non-Manual features for subunit level medical sign modelling in automatic sign language classification and recognition”. In: *Journal of Medical Systems* 11 (Sept. 2017). ISSN: 1573-689X.

[37] R. Elliott et al. “An overview of the SiGML notation and SiGML signing software system”. English. In: *Sign Language Processing Satellite Workshop of the Fourth International Conference on LREC*. May 2004, pp. 98–104.

[38] S. Escalera et al. “ChaLearn looking at people challenge 2014: Dataset and results”. In: *ECCV Workshops (1)*. 2014, pp. 459–473.

[39] S. Escalera et al. “Multi-modal gesture recognition challenge 2013: Dataset and results”. In: *Proc. of the 15th ACM on International Conference on Multimodal Interaction (ICMI)*. ACM. 2013, pp. 445–452.

[40] J. Frankel et al. “Articulatory feature classifiers trained on 2000 hours of telephone speech”. In: *Proc. of Interspeech*. 2007.

[41] N. Gamage et al. “Gaussian process dynamical models for hand gesture interpretation in sign language”. In: *Pattern Recognition Letters* 32 (2011), pp. 2009–2014.

[42] F. Gaolin et al. “A novel approach to automatically extracting basic units from Chinese Sign Language”. In: *Proc. of the 17th International Conference on Pattern Recognition*. Aug. 2004.

[43] A. Graves et al. “Connectionist Temporal Classification: Labelling unsegmented sequence data with recurrent neural networks”. In: *Proc. of the ACM International Conference on Machine Learning (ICML)*. 2006.

[44] J. Han, G. Awad, and A. Sutherland. “Boosted subunits: a framework for recognising sign language from videos”. In: *IET Image Processing* 7.1 (Feb. 2013). ISSN: 1751-9659.

[45] T. Hanke. “HamNoSys - Representing sign language data in language resources and language processing contexts”. In: *Workshop proceedings : Representation and processing of sign languages* (2004), pp. 1–6.

[46] H. Hermansky. “Perceptual Linear Predictive (PLP) analysis of speech”. In: *Journal of the Acoustical Society of America* 57.4 (Apr. 1990), pp. 1738–52.

[47] H. Hermansky, D. Ellis, and S. Sharma. “Tandem connectionist feature extraction for conventional HMM systems”. In: *Proc. of the IEEE ICASSP*. Vol. 3. 2000, pp. 1635–1638.

[48] H. Hienz, B. Bauer, and K.-F. Kraiss. “HMM-Based continuous sign language recognition using stochastic grammars”. In: *Gesture-Based Communication in Human-Computer Interaction*. Springer Berlin Heidelberg, 1999, pp. 185–196. ISBN: 978-3-540-46616-1.

[49] ISARA application. *ISARA application*. URL: https://isara.app (visited on 02/27/2021).

[50] A. Jansen, K. Church, and H. Hermansky. “Towards spoken term discovery at scale with zero resources”. In: *Proc. of Interspeech*. 2010, pp. 1676–1679.

[51] F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press 1998, pp. 65, 70, 71, 73, 74. ISBN: 0262-10066-5.

[52] D. Johnson et al. *ICSI Quicknet Software Package*. 2004. URL: http://www.icsi.berkeley.edu/Speech/qn.html (visited on 02/27/2021).

[53] H. Junwei, A. George, and S. Alistair. "Modelling and segmenting subunits for sign language recognition based on hand motion analysis". In: *Pattern Recognition Letters* 30.6 (2009), pp. 623–633. ISSN: 0167-8655.

[54] T. Kadir et al. "Minimal training, large lexicon, unconstrained sign language recognition". In: *Proc. of the British Machine Vision Conference (BMVC)*. Vol. 2. 2004, pp. 939–948.

[55] T. Kailath. "The divergence and Bhattacharyya distance measures in signal selection". In: *IEEE Transactions on Communication Technology* 15.1 (Feb. 1967), pp. 52–60. ISSN: 0018-9332.

[56] R.M. Kaplan and M. Kay. "Regular models of phonological rule systems". In: *Computational Linguistics* 20 (1994), pp. 331–378.

[57] C. Keskin and L. Akarun. "STARS: Sign tracking and recognition system using input–output HMMs". In: *Pattern Recognition Letters* 30.12 (2009). Image/video-based Pattern Analysis and {HCI} Applications, pp. 1086–1095. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2009.03.016. URL: http://www.sciencedirect.com/science/article/pii/S0167865509000543.

[58] S. King et al. "Speech production knowledge in automatic speech recognition". In: *Journal of the Acoustical Society of America* 121.2 (Feb. 2007), pp. 723–742.

[59] O. Koller, H. Ney, and R. Bowden. "Automatic alignment of HamNoSys subunits for continuous sign language recognition". In: *Proc. of the Tenth International Conference on LREC*. 2016.

[60] O. Koller, H. Ney, and R. Bowden. "Deep Hand: How to train a CNN on 1 Million hand images when your data is continuous and weakly labelled". In: *Proc. of the IEEE CVPR*. June 2016.

[61] O. Koller, H. Ney, and R. Bowden. "May the force be with you: Force-aligned signwriting for automatic subunit annotation of corpora". In: *Proc. of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (AFGR)*. Apr. 2013.

[62] O. Koller et al. "Deep sign: hybrid CNN-HMM for continuous sign language recognition". In: *Proc. of the BMVC*. 2016.

[63] W. Kong and S. Ranganath. "Towards subject independent continuous sign language recognition: A segment and merge approach". In: *Pattern Recognition* 47 (2014), pp. 1294–1308.

[64] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Proc. of the Advances in Neural Information Processing Systems (NIPS)*. 2012.

[65] S. Kullback and R. A. Leibler. "On information and sufficiency". In: *The Annals of Mathematical Statistics* (1951). DOI: 10.1214/aoms/1177729694. URL: https://doi.org/10.1214/aoms/1177729694.

[66] P. Kumar et al. "Coupled HMM-based multi-sensor data fusion for sign language recognition". In: *Pattern Recognition Letters* 86 (2017), pp. 1–8. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2016.12.004. URL: http://www.sciencedirect.com/science/article/pii/S0167865516303518.

[67] V. I. Levenshtein. "Binary codes capable of correcting deletions, insertions, and reversals". In: *Soviet Physics Doklady* 10.8 (Feb. 1966), pp. 707–710.

[68] T. H. S. Li, M. C. Kao, and P. H. Kuo. "Recognition system for home-service-related sign language using entropy-based k-means algorithm and ABC-based HMM". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 46.1 (Jan. 2016), pp. 150–162. ISSN: 2168-2216. DOI: 10.1109/TSMC.2015.2435702.

[69] J. Lichtenauer, E. Hendriks, and M. Reinders. "Learning to recognize a sign from a single example". In: *Proc. of the 8th IEEE International Conference on Automatic Face and Gesture Recognition (AFGR)*. Sept. 2008, pp. 1–6. DOI: 10.1109/AFGR.2008.4813450.

[70] J.F. Lichtenauer, E.A. Hendriks, and M.J.T. Reinders. "Sign language recognition by combining statistical dtw and independent classification". In: *IEEE Transactions on PAMI* 30 (2008), pp. 2040–2046.

[71] S. K. Liddell and R. E. Johnson. "American Sign Language: The phonological base ". In: *Sign Language Studies*. Vol. 64. 1989, pp. 195–277.

[72] N. Liu et al. "Model structure selection training algorithms for an HMM gesture recognition system". In: *Ninth International Workshop on Frontiers in Handwriting Recognition*. 2004, pp. 100–105.

[73] K. Livescu et al. "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU Summer Workshop". In: *Proc. of the IEEE ICASSP*. 2007.

[74] T. Matsuo, Y. Shirai, and N. Shimada. "Automatic generation of HMM topology for sign language recognition". In: *2008 19th International Conference on Pattern Recognition*. Dec. 2008, pp. 1–4. DOI: 10.1109/ICPR.2008.4761525.

[75] G. Morgan and B. Woll. "The development of complex sentences in British Sign Language". In: *Directions in Sign Language Acquisition: Trends in Language Acquisition Research*. Ed. by Gary Morgan et al. John Benjamins, Amsterdam, Netherlands, 2002, pp. 255–276.

[76] K. Murakami and H. Taguchi. “Gesture recognition using recurrent neural networks”. In: *Procs. of SIGCHI Conf. on Human factors in computing systems: Reaching through technology*. ACM New York. NY, USA, 1991, pp. 237–242.

[77] V. Nyst. *24. Shared sign languages*. Berlin, Boston: De Gruyter Mouton, 16 Aug. 2012, pp. 552–574. ISBN: 9783110204216. DOI: https://doi.org/10.1515/9783110261325.552. URL: https://www.degruyter.com/view/book/9783110261325/10.1515/9783110261325.552.xml (visited on 02/27/2021).

[78] M. Oliveira et al. “Irish Sign Language recognition using principal component analysis and convolutional neural networks”. In: *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. 2017, pp. 1–8. DOI: 10.1109/DICTA.2017.8227451.

[79] E.-J. Ong et al. “Sign language recognition using sequential pattern trees”. In: *Proc. of the IEEE CVPR*. 2012.

[80] S. C. W. Ong and S. Ranganath. “Automatic sign language analysis: A survey and the future beyond lexical meaning.” In: *IEEE Transactions on PAMI* 27.6 (2005), pp. 873–891.

[81] OpenNI organization. *OpenNI user guide*. Cambridge University Engineering Department, 2010.

[82] V. Pagel, K. Lenzo, and A.W. Black. “Letter to sound rules for accented lexicon compression”. In: *Proc. of International Conference on Spoken Language Processing*. 1998.

[83] A. S. Park and J. R. Glass. “Towards unsupervised pattern discovery in speech”. In: *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* 2005, pp. 53–58.

[84] A. S. Park and J. R. Glass. “Unsupervised pattern discovery in speech”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 16.1 (2008), pp. 186–197.

[85] V. N. Pashaloudi and K. G. Margaritis. “Hidden Markov model for sign language recognition: A review”. In: *Proc. 2nd Hellenic Conf. AI, SETN-2002*. 2002, pp. 11–12.

[86] A. Paszke et al. “PyTorch: An imperative style, high-performance deep learning library”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 8024–8035.

[87] E. Pignat and S. Calinon. “Learning adaptive dressing assistance from human demonstration”. In: *Robotics and Autonomous Systems* 93 (July 2017), pp. 61–75.

[88] L. Pigou et al. “Sign language recognition using convolutional neural networks”. In: *Computer Vision - ECCV 2014 Workshops*. Ed. by Lourdes Agapito, Michael M. Bronstein, and Carsten Rother. Cham: Springer International Publishing, 2015, pp. 572–578. ISBN: 978-3-319-16178-5.

[89] J. F. Pitrelli et al. “PhoneBook: a phonetically-rich isolated-word telephone-speech database”. In: *Proc. of the ICASSP*. Vol. 1. May 1995, 101–104 vol.1. DOI: 10.1109/ICASSP.1995.479283.

[90] V. Pitsikalis et al. “Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition”. In: *Proc in the IEEE CVPR Workshops*. June 2011.

[91] Y. Quan and P. Jinye. “Chinese Sign Language recognition for a vision-based multi-features classifier”. In: *International Symposium on Computer Science and Computational Technology*. Shanghai, China, pp. 194–197.

[92] L. R. Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition”. In: *Proc. of the IEEE* 77.2 (1989), pp. 257–286.

[93] R. Rasipuram and M. Magimai.-Doss. “Acoustic and lexical resource constrained ASR using language-independent acoustic model and language-dependent probabilistic lexical model”. In: *Speech Communication* 68 (Apr. 2015), pp. 23–40.

[94] R. Rasipuram and M. Magimai.-Doss. “Articulatory feature based continuous speech recognition using probabilistic lexical modeling”. In: *Computer Speech and Language* 36 (2016), pp. 233–259. ISSN: 0885-2308. DOI: 10.1016/j.csl.2015.04.003.

[95] M. Razavi, R. Rasipuram, and M. Magimai.-Doss. “Acoustic data-driven grapheme-to-phoneme conversion in the probabilistic lexical modeling framework”. In: *Speech Communication* 80 (2016).

[96] R.-H. Liang and M. Ouhyoung. “A real-time continuous gesture recognition system for sign language”. In: *Proc. Third IEEE International Conference on Automatic Face and Gesture Recognition*. 1998, pp. 558–567.

[97] M. I. Sadek, M. N. Mikhael, and H. A. Mansour. “A new approach for designing a smart glove for Arabic Sign Language recognition system based on the statistical analysis of the sign language”. In: *2017 34th National Radio Science Conference (NRSC)*. 2017, pp. 380–388.

[98] S. Sako and T. Kitamura. “Subunit modeling for Japanese Sign Language recognition based on phonetically depend multi-stream hidden Markov models”. In: *Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion*. Springer Berlin Heidelberg, 2013, pp. 548–555. ISBN: 978-3-642-39188-0.

[99] J. Segen and S. Kumar. “Shadow gestures: 3D hand pose estimation using a single camera”. In: *Proc. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*. Vol. 1. 1999, 479–485 Vol. 1.

[100] J. Shotton et al. “Real-time human pose recognition in parts from single depth images”. In: *CVPR 2011*. 2011, pp. 1297–1304.

[101] S. Siddiqi, G. Gordon, and A. Moore. “Fast state discovery for HMM model selection and learning”. In: *Proc. of the Eleventh International Conference on Artificial Intelligence and Statistics (AI-STATS)*. 2007.

[102] T. Starner. “Visual recognition of American Sign Language using hidden Markov models”. PhD thesis. Cambridge, Mass: Media Laboratory, Massachusetts Institute of Technology, 1995.

[103] T. Starner and A. Pentland. “Real-time American Sign Language recognition from video using hidden Markov models”. In: *Proc. of International Symposium on Computer Vision - ISCV*. 1995.

[104] T. Starner, J. Weaver, and A. Pentland. “Real-time American Sign Language recognition using desk and wearable computer based video”. In: *IEEE Transactions on PAMI* 20.12 (1998), pp. 1371–1375.

[105] W. Stokoe. “An outline of the visual communication systems of the American deaf”. In: *Studies in linguistics: Occasional papers* 86 (1960).

[106] S. Stoll, S. Hadfield, and R. Bowden. “SignSynth: Data-Driven Sign Language Video Generation”. In: *Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*. Ed. by Adrien Bartoli and Andrea Fusiello. Vol. 12538. Lecture Notes in Computer Science. Springer, 2020, pp. 353–370.

[107] S. Stoll et al. “Text2Sign: Towards Sign Language Production using Neural Machine Translation and Generative Adversarial Networks”. In: (2019), pp. 1–18.

[108] H. Strik and C. Cucchiarini. “Modeling pronunciation variation for ASR: A survey of the literature”. In: *Speech Communication* 29.2-4 (July 14, 2009), pp. 225–246.

[109] R. Sutton-Spence and B. Woll. *The linguistics of British Sign Language: An introduction*. Cambridge University Press, 1999. DOI: 10.1017/CBO9781139167048.

[110] M. N. Suzgun et al. “HOSPISIGN: An interactive sign language platform for hearing impaired”. In: *Journal of Naval Science and Engineering* 11.3 (2015), pp. 75–92.

[111] S. Theodorakis, V. Pitsikalis, and P. Maragos. “Model-level data-driven sub-units for signs in videos of continuous sign language”. In: *Proc. in the IEEE ICASSP*. Mar. 2010, pp. 2262–2265.

[112] SignAll Technologies Inc. (USA). *SignAll*. URL: https://www.signall.us/ (visited on 02/27/2021).

[113] B. Varadarajan, S. Khudanpur, and E. Dupoux. “Unsupervised learning of acoustic sub-word units”. In: *Proc. of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. 2008, pp. 165–168.

[114] M. Versteegh et al. “The zero resource speech challenge 2015”. In: *Proc. of Interspeech*. 2015.

[115] C. Vogler and D. Metaxas. “Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods”. In: *IEEE International Conference on Systems, Man and Cybernetics (SMC)*. 1997, pp. 156–161.

[116] C. Vogler and D. Metaxas. “ASL recognition based on a coupling between HMMs and 3D motion analysis”. In: *Sixth International Conference on Computer Vision (ICCV ’98)*. Washington, DC, USA: IEEE Computer Society, 1998, p. 363. ISBN: 81-7319-221-9.

[117] C. Vogler and D. Metaxas. “Parallel hidden Markov models for American Sign Language recognition”. In: *Proc. of the Seventh IEEE ICCV*. Vol. 1. Sept. 1999, 116–122 vol.1. DOI: 10.1109/ICCV.1999.791206.

[118] M.B. Waldron and S. Kim. “Isolated ASL sign recognition system for deaf persons”. In: *IEEE Transactions on Rehabilitation Engineering* 3(3) (1995), pp. 261–271.

[119] D. Wang and S. King. “Letter-to-sound pronunciation prediction using conditional random fields”. In: *Signal Processing Letters, IEEE* 18.2 (2011), pp. 122–125.

[120] H. Wang et al. “Fast sign language recognition benefited from low rank approximation”. In: *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 1. hmm model selection. May 2015, pp. 1–6. DOI: 10.1109/FG.2015.7163092.

[121] P. Wang et al. “RGB-D-based human motion recognition with deep learning: A survey”. In: *Computer Vision and Image Understanding* 171 (2018), pp. 118–139. ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2018.04.007. URL: https://www.sciencedirect.com/science/article/pii/S1077314218300663.

[122] S.-F. Wong and R. Cipolla. “Real-time interpretation of hand motions using a sparse Bayesian classifier on motion gradient orientation images”. In: *Proc. of the British Machine Vision Conference (BMVC)*. Vol. 1. Oxford, UK, Sept. 2005, pp. 379–388.

[123] D. Wu et al. “Deep dynamic neural networks for multimodal gesture segmentation and recognition”. In: *IEEE transactions on PAMI* 38.8 (2016), pp. 1583–1597.

[124] Q. Xiao et al. “Multimodal Fusion Based on LSTM and a Couple Conditional Hidden Markov Model for Chinese Sign Language Recognition”. In: *IEEE Access* 7 (2019), pp. 112258–112268. DOI: 10.1109/ACCESS.2019.2925654.

[125] C. Xie, L. Yu, and S. Wang. “Deep feature extraction and multi-feature fusion for similar hand gesture recognition”. In: *2018 IEEE Visual Communications and Image Processing (VCIP)*. 2018, pp. 1–4. DOI: 10.1109/VCIP.2018.8698688.

[126] S. Xie et al. “Aggregated residual transformations for deep neural networks”. In: *2017 IEEE Conference on CVPR*. 2017, pp. 5987–5995. DOI: 10.1109/CVPR.2017.634.

[127] M.H. Yang, N. Ahuja, and M. Tabb. “Extraction of 2D motion trajectories and its application to hand gesture recognition”. In: *IEEE TPAMI* 24 (2002), pp. 1061–1074.

[128] W. Yang, J. Tao, and Z. Ye. “Continuous sign language recognition using level building based on fast hidden Markov model”. In: *Pattern Recognition Letters* 78 (2016), pp. 28–35. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2016.03.030. URL: http://www.sciencedirect.com/science/article/pii/S0167865516300344.

[129] A. Yilmaz, O. Javed, and M. Shah. “Object tracking: A survey”. In: *ACM Comput. Surv.* 38.4 (Dec. 2006), 13–es. ISSN: 0360-0300. DOI: 10.1145/1177352.1177355. URL: https://doi.org/10.1145/1177352.1177355.

[130] S. Young et al. *The HTK Book Version 3.0*. Cambridge University Press, 2000.

[131] M. M. Zaki and S. I. Shaheen. “Sign language recognition using a combination of new vision based features”. In: *Pattern Recognition Letters* 32.4 (2011), pp. 572–577. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2010.11.013. URL: http://www.sciencedirect.com/science/article/pii/S016786551000379X.

[132] Z. Zhang. “Microsoft Kinect sensor and its effect”. In: *IEEE MultiMedia* 19.2 (2012), pp. 4–10.

SANDRINE TORNAY

RESEARCH ASSISTANT - PHD STUDENT

ABOUT ME

I like to resolve problems; designed, developed and applied in all fields. I am interested in any challenge, if it can help others it is a big plus. I am an organized, creative, thoughtful person who is curious to learn.

ACTUALLY

2016 - (01.2021 expecting)

PhD in Electrical Engineering - PHONOLOGY-BASED SIGN LANGUAGE RECOGNITION AND ASSESSMENT

Ecole polytechnique fédérale de Lausanne, Idiap Research Institute (Martigny)

2019 - ...

Graphic Designer Student

Design et formations (online school)

CONTACT

079 236 42 34

sandrine.tornay@gmail.com

Rue des Sondzons 33
1904 Vernayaz

DIPLOMAS

MASTER of Sciences
MATHEMATICS
University of Fribourg, 2015

BACHELOR of Sciences
MATHEMATICS, Psychology, Neurosciences
University of Fribourg, 2014

CIVIL STATUS

BIRTH	22.08.1989
NATIONALITY	Switzerland
CIVIL STATUS	Married
MAIDEN NAME	Revaz
CHILDREN	1

WORK EXPERIENCES

2016 - ...

Idiap Research Institut

Rue Marconi 19, CH-1920 Martigny

RESEARCH ASSISTANT

My research focuses on how to model isolated signs of sign languages for recognition and assessment tasks. I am involved in the multidisciplinary framework of the SMILE project for which we have developped a demo. Here are presentation videos:

French: *English:*

2015 - 2016 (30%)

Ecole Professionnelle Technique et des Métiers Sion

Chemin Saint-Hubert 2, CH-1950 Sion

MATHEMATICS TEACHER

I taught mathematics (geometry, analysis, linear algebra) to apprentices who made the professional maturity in parallel with their CFC (students from 16 to 20 years old). The teaching program was fixed in advance (parallel teaching).

03.2016 - 05.2016 (60%)

Icare Institut

Rue de Technopôle 10, CH-3960 Sierre

TRAINEE

I have developed a program that predicts daily energy consumption. I used recurrent neural network (LSTM).

SKILLS

- Scientific research
- Critical Thinking
- Problem solving
- Programming (Python, bash)
- Demo development
- Scientific writing
- Multidisciplinary collaboration
- Public presentation
- Scientific presentation
- Good time management

WORK EXPERIENCES (CONTINUED)

2014

Idiap Research Institut

Rue Marconi 19, CH-1920 Martigny

MASTER THESIS

My Master thesis presents mathematical methods encountered in automatic speech recognition (ASR). To complete the theory part, experiments were done at Idiap Research Institut using methods based on hidden Markov models.

2011 - 2013

University of Fribourg

Av. de l'Europe 20, CH-1700 Fribourg

PROPAEDEUTIC ANALYSIS TUTOR

I gave propaedeutic analysis exercise sessions to students doing a Bachelor of sciences (around 200 students). More specifically, exercise presentation sessions and correction sessions. I also participated in the correction of the series of exercises.

2010 - 2012

Banque Raiffeisen Martigny
DepioPharm Martigny
Ravoire Colony
Grand Dixence Hotel

VARIOUS SUMMER WORKS

I did various summer works during my school career: administrative work, support course teacher (bachelor level, primary and secondary school level), child care (au pair stay in England), colony monitor, various jobs in the hotel industry.

SKILLS

- Do state-of-the-art
- Teaching (Math)
- Lead an exercise session
- Course organisation
- Events organisation
- Autonomy
- Adaptability
- Curiosity
- Empathy

PUBLISHED PAPERS *available on http://publications.idiap.ch/*

Subunits Inference and Lexicon Development Based on Pairwise Comparison of Utterances and Signs

Sandrine Tornay and Mathew Magimai.-Doss, in: Information, 10:298, 2019, https://www.mdpi.com/2078-2489/10/10/298

HMM-based Approaches to Model Multichannel Information in Sign Language inspired from Articulatory Features-based Speech Processing

Sandrine Tornay, Marzieh Razavi, Necati Cihan Camgoz, Richard Bowden and Mathew Magimai.-Doss,
in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019

Towards Multilingual Sign Language Recognition

Sandrine Tornay, Marzieh Razavi and Mathew Magimai.-Doss, in: Proc. of IEEE ICASSP, 2020

An HMM Approach with Inherent Model Selection for Sign Language and Gesture Recognition

Sandrine Tornay, Oya Aaran and Mathew Magimai.-Doss, in: Language Ressources and Evaluation Conference (LREC), 2020

A Phonology-based Approach for Isolated Sign Production Assessment in Sign Language

Sandrine Tornay, Necati Cihan Camgoz, Richard Bowden and Mathew Magimai.-Doss, in: International Conference on Multimodal Interaction, 2020

EXTRA

Cashier, Secretary

Cave de la Grand'Rue, CH-1904 Vernayaz

Business Concept Certificat

Innosuisse Start-up Training,
EPFL Innovation Park, CH-1015 Lausanne

LANGUAGES

- French (Mother tongue)
- English (B2 level)
- German (school level)
- French Sign Language (A1 level)

HOBBIES

- Intuitive painting
- Flute
- Manual conception
  (sewing, knitting, scrapbooking, ...)
- Meditation, Self-development
- Event organisation
- Local politic
- Renovation

PROFESSIONAL REFERENCE

Main thesis supervisor

MATHEW MAGIMAI.-DOSS
Senior Researcher
Idiap Research Institute

+4127 721 77 88

mathew@idiap.ch