Modelling the Distribution of Human Motion for Sign Language Assessment

Oliver Cory[1], Ozge Mercanoglu Sincan[1], Matthew Vowels[1,2,3], Alessia Battisti[4], Franz Holzknecht[5], Katja Tissi[5], Sandra Sidler-Miserez[5], Tobias Haug[5], Sarah Ebling[4], and Richard Bowden[1]

[1] Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK
{o.cory,o.mercanoglusincan,r.bowden,m.j.vowels}@surrey.ac.uk
[2] The Sense Innovation and Research Center, Lausanne and Sion, CH
[3] Institute of Psychology, University of Lausanne (UNIL), Lausanne, CH
[4] Department of Computational Linguistics, University of Zurich, Zurich, CH
{alessia.battisti,sarah.ebling}@uzh.ch
[5] University of Teacher Education in Special Needs (HfH), Zurich, CH
{franz.holzknecht,katja.tissi}@hfh.ch, sandysidler@gmail.com

Abstract. Sign Language Assessment (SLA) tools are useful to aid in language learning and are underdeveloped. Previous work has focused on isolated signs or comparison against a single reference video to assess Sign Languages (SL). This paper introduces a novel SLA tool designed to evaluate the comprehensibility of SL by modelling the natural distribution of human motion. We train our pipeline on data from native signers and evaluate it using SL learners. We compare our results to ratings from a human raters study and find strong correlation between human ratings and our tool. We visually demonstrate our tools ability to detect anomalous results spatio-temporally, providing actionable feedback to aid in SL learning and assessment.

Keywords: Sign Language Assessment · Human Motion Modelling

1 Introduction

Sign Languages (SL) are nuanced and complex visual-gestural languages that are the primary form of communication for millions of deaf [6] people worldwide. With the advancements in deep learning and computer vision, there has been a growing interest in modelling SL. The majority of methods focus on classification, namely for the recognition and translation of sign [4,30], rather than improving or assessing SL proficiency. The standardisation of Sign Language Assessment (SLA) is a challenging research topic due to the many nuances that affect its legibility [14].

---

[6] We follow the recent convention of abandoning a distinction between *Deaf* and *deaf* and use the latter term also to refer to (deaf) members of the sign language community [33,38].

The study of Sign Language Linguistics is still in its infancy, especially when compared to spoken languages. SL have no standardised written form, they are conveyed via a combination of manual and non-manual features [41]. While the manual features include the location, orientation, and movement of the arms and hands; non-manual features refer to facial expressions, body posture, head movement, and eye gaze. Signing involves simultaneous combinations of these features, each influencing the meaning of a sign, adding multiple layers of linguistic complexity. In continuous sequences, co-articulation is also common factor [42]. This includes temporal overlap between signs in a sequence leading to blending, spatial influence where the location of one sign may impact the starting location of the following signs, and handshape modifications based on context. Given the rich and complex nature of SL, skilled teachers are needed to assess and quantify signing proficiency.

In this paper we focus on SL assessment, proposing a tool to aid human teachers to evaluate continuous SL and to improve efficiency in evaluation and feedback. Teaching systems for SL that incorporate feedback mechanisms have been proposed using classification to determine correct from incorrect repetitions or to regress scores directly [49,54]. However, most approaches are limited to the assessment of isolated signs [50].

Our work provides an SL assessment tool for continuous sequences that learns the natural distribution present in human motion. We develop a *Skeleton Variational Autoencoder (SkeletonVAE)* to embed signed sequences from multiple native signers in a compact, lower dimensional subspace. We then apply a *Reference Selection* technique over these embeddings to determine the most representative sequence from the collection of sequences. We finally model the *Motion Envelope* by aligning all the sequences to the reference and learning the distribution over the embedded data using a Gaussian Process (GP).

We test our model using data from SL learners and evaluate its performance against ratings collected from a human raters study. We demonstrate that our model can quantitatively evaluate the production of sequences achieving similar results to a manual rater. Furthermore, we show that our system can determine where and by what distance a learner falls outside of the natural acceptable variation in human motion for signed sequences.

2   Related Work

Sign Language Recognition, Translation and Production. Computational approaches to SL modelling have been the focus of researchers for over 30 years [47]. Preliminary research focused on isolated Sign Language Recognition (SLR) using statistical methods [12], aiming to classify isolated signs. The advent of deep learning techniques has enabled the development of continuous SLR methods operating over continuous sequences, implemented using CNNs [29,30], RNNs [6,31] and more recently Transformers [5]. Some researchers operate in the pixel space directly whereas others choose to use skeleton pose, optical flow, or a combination of modalities [22,44].

More recently the field has moved towards Sign Language Translation (SLT) [4], aiming to translate continuous sign to written spoken language sentences rather than just recognising the consistent signs. SLT is a more challenging task than SLR due to the grammatical and ordering differences between SL and spoken language. Transformer-based approaches have achieved state-of-the-art performance, learning the recognition and translation tasks jointly [5].

Most SLT approaches require intermediate representations and while traditional approaches often rely on linguistic representations such as gloss [4, 5] or HamNoSys [28, 53], such annotation is expensive to create. To overcome this bottleneck, recent research has shifted towards gloss-free translation [11, 52, 55, 58, 59].

On the other hand, Sign Language Production (SLP) aims to produce SL videos from written spoken language sentences. Current approaches to SLP use Transformer-based architectures, extending SLT to include the production of digital avatars [23], photo-realistic outputs using Generative Adversarial Networks [43, 46] or diffusion models [8].

Language Learning and Assessment. Automated tools for language learning have been widely developed for written and spoken languages. Mainstream tools such as Duolingo [7] have proven their effectiveness in increasing learning efficiency through gamification. There are only a few studies that utilize gamification for SL which aim to teach isolated signs [2, 45].

Sign Language Assessment (SLA) systems that provide more detailed feedback have also been introduced [48–50]. Tornay et al. [50] provide a scoring mechanism alongside a visualisation showing the performed sign against a reference skeleton, providing actionable feedback. However, this is limited to isolated signs. Wen et al. [54] introduced an approach for SL assessment over continuous sequences using a two-stage method that integrated domain knowledge from action similarity techniques. However, the method relies on a single reference video to evaluate against and therefore does not account for the natural variation in human motion during assessment.

The complexity of SL makes its annotation and assessment challenging. Annotating SL data is extremely time consuming, with one minute of sign taking between 10 and 30 minutes to annotate [24]. The natural variability in signing between individuals (often referred to as signer 'style' [32]) further complicates data annotation and quantification of SL proficiency. Human assessment remains the most reliable method for scoring SL, as teachers can accurately determine correct sign production despite natural variation between signers [1]. Holzknecht et al. [18] compare the results of an automated SLA system with ratings from a human rater study for isolated signs. We compare our approach to human raters in the context of continuous signed sequences.

Action Quality Assessment. The assessment of SL can be seen as a subdomain within the broader field of Action Quality Assessment (AQA), which aims to evaluate the quality and performance of human actions in various contexts.

Previous research in AQA has typically focused on macro-level actions [21,36,56] rather than fine details like individual hand and finger movements. The majority of these methods compare actions against a single reference and directly regress a score [21,56]. Some use additional hardware to track human behaviours [34], some work directly from RGB videos [56], and others use pose representations [36,54].

Morais et al. [36] employ a pose representation for anomaly detection. However, this method requires high correspondence of movements to be effective, with performance degrading due to the natural variations in posture between individuals. Xu et al. [56] rely on meticulously annotated training data to achieve high accuracy results , making it a highly supervised approach.

Two-stage approaches have emerged offering a more flexible and interpretable framework for assessing action quality by decoupling feature extraction from the evaluation process, allowing for more adaptable and insightful AQA systems [9, 20, 54]. In the domain of SL, Wen et al. [54] proposes a two-stage pipeline where features are first recovered from video, embedded and aligned to a single reference using Dynamic Time Warping [40] before assessing sign quality.

Distinct from previous approaches, our unsupervised method accounts for the natural variation in human motion when assessing action quality by learning the distribution in motion over multiple expert productions.
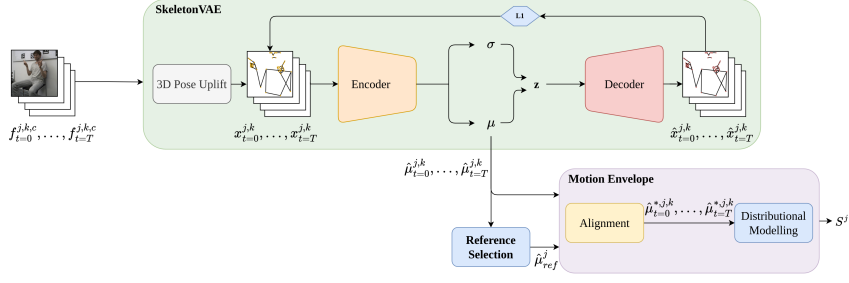
3   Method

We present a novel approach for learning the natural distribution of continuous Sign Language sequences. We first build a *SkeletonVAE* by uplifting multi-view video data to a 3D skeleton pose and learning a low-dimensional latent representation of pose, capturing the essential characteristics of human movement. We take our video dataset of sentences with multi-participant productions and encode them to create a secondary dataset of latent time varying embeddings. Second, we develop a *Reference Selection* technique which identifies a reference production of each sentence based on similarity calculation between all participants. Finally, we build a *Motion Envelope* by aligning each participant's sequence to its corresponding sentence reference and model the distribution of per-dimension embedding trajectories across multiple signers. The pipeline for this method is shown in Fig. 1.

3.1   SkeletonVAE

Consider N sequences of SL video frames $\mathbf{f}_t^{j,k,c}$, where $j = \{1, 2..., J\}$ sign language sentences being executed by $k = \{1, 2, ..., K\}$ individual signers, and where t is an individual timepoint ranging $t = \{1, 2, ..., T_{j,k}\}$, such that the total number of timepoints depends on the signer and the sentence being performed. c indexes C synchronised cameras that capture all the data together.

We start by extracting Mediapipe [35] 2D poses from a single view for C cameras over the entire dataset. After this, we implement 3D pose uplift [19] to regress accurate 3D skeleton data and convert to canonical form by choosing

Fig. 1: Diagram showing training pipeline for modelling the jth sentence over K signers. The process takes J example sentences captured with C independent cameras and uses 3D pose uplift to create a set of $\mathbf{x}$ poses which are fed into the VAE, encoding the poses into $\hat{\boldsymbol{\mu}}$ latent means. Reference Selection finds the central signal $\hat{\boldsymbol{\mu}}_{ref}$ and learns a distribution over K signers.

fixed bone lengths and applying this scaling via the joint angles. We now have N sequences of d-dimensional skeleton joint-position data for sign language poses $\mathbf{x}_t^{j,k}$.

We assume that, within the context of human motion and SL, each pose lies on some manifold with fewer dimensions than d which we can approximate via a stochastic mapping $p_\theta(\mathbf{z}|\mathbf{x}_t^{j,k}) : \mathbf{x} \to \mathbf{z}$ where $\mathbf{z} \in \mathbb{R}^\Omega$ is a latent representation or embedding. Our goal is to model the time variation of $\mathbf{x}$ in terms of its compact representation $\mathbf{z}$.

We begin by taking the skeleton poses $\mathbf{x}$ and embedding them using a Variational AutoEncoder, which is trained via a process known as variational inference [3,16,26]. Variational inference is concerned with maximising the Evidence Lower BOund (ELBO), which forms a lower bound on the negative log-likelihood of the data under the model:

$$N^{-1} \sum_{i=1}^{N} \log p_\theta(\mathbf{x_i}) \leq$$

$$N^{-1} \sum_{i}^{N} \left(-\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \left[\log p_\theta(\mathbf{x}_i|\mathbf{z}) \right] + \beta \mathbb{D}_{KL} \left[q_\phi(\mathbf{z}|\mathbf{x}_i) || p(\mathbf{z}) \right] \right) \ . \tag{1}$$

Here, $q_\phi(\mathbf{z}|\mathbf{x})$ is known as the approximating posterior, which ideally matches the true posterior $p(\mathbf{z}|\mathbf{x})$ which we do not have access to. We therefore assume a parameterisation for this approximating posterior, and define a prior distribution $p(\mathbf{z})$. The Kullback-Liebler divergence $\mathbb{D}_{KL}$ is then used to create pressure such that the approximating posterior distribution q resembles this prior, and this pressure is weighted with a scalar β [16]. Using a β value other than one means the ELBO cannot technically be fulfilled, but is a hyperparameter determined by experimental results. For our work the choice of prior is an isotropic Gaussian with mean $\boldsymbol{\mu} = 0$ and variance $\boldsymbol{\sigma}^2 = 1$. The parameters ϕ and θ represent the

neural network parameters for the encoder and decoder respectively, and it is the encoder and decoder which parameterise the approximating posterior and conditional likelihood models. As such, each datapoint is encoded as a mean $\hat{\boldsymbol{\mu}}$ and a variance $\hat{\boldsymbol{\sigma}}^2$ which, via the reparameterisation trick [27], enable sampling $\mathbf{z}|\mathbf{x}_i \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\sigma}}_i^2)$ which are decoded to reconstructions $\hat{\mathbf{x}}_i$. Finally, note that in Eq. 1, N is the total number of datapoints available, across all J sentences and K signers (i.e. $N = K \times J \times T$ for fixed T). This part of the modeling process therefore treats the data as independent and identically distributed (the sequential aspect of the data, as well as the fact we have different sentences being performed, will be modeled using Gaussian Processes).

$$\mathcal{L}_{\mathbf{VAE}} = \alpha \mathbf{L1}_{\mathbf{hands}} + (1 - \alpha)\mathbf{L1}_{\mathbf{body}} + \beta \mathbb{D}_{\mathbf{KL}} \tag{2}$$

Since hands are high-frequency, low-amplitude signals due to their rapid and detailed movements compared to the larger, slower movements of the body, they can be lost in the noise during VAE training. To address this, we use L1 loss as the reconstruction loss and split the weighting of the loss between the hands and body. By setting a high α value, the network can better focus on hand reconstruction. Our overall loss function, Eq. (2), is the sum of this reconstruction loss with the β-scaled KLD. Once we have trained the VAE on all skeleton poses for the complete dataset, we arrive at a secondary dataset of encodings $\hat{\boldsymbol{\mu}}_t^{j,k}$ where $\hat{\boldsymbol{\mu}}$ represents the conditional mean encoding of the corresponding skeleton datapoint $\mathbf{x}$.

3.2   Reference Selection

For J sentences, we calculate a cosine similarity matrix comparing the encoded means $\hat{\boldsymbol{\mu}}$ over K signers. We then average the matrix entries for each k, returning the average similarity scores of k with reference to all other k's that produced j. We choose the highest average similarity signal as our reference signal $\hat{\boldsymbol{\mu}}_{ref}^j$, for each sentence. This signal is the central signal and is used as the reference for the Dynamic Time Warping (DTW) [40] algorithm.

3.3   Motion Envelope

At this stage we use DTW to align the sequences such that $T_{j,k} = T_j^* \forall k$, where T_j^* is the length of the reference sequence $\hat{\boldsymbol{\mu}}_{ref}^j$. Each of the aligned sequences are denoted $\hat{\boldsymbol{\mu}}^*$.

Finally we train a Gaussian Process (GP) [39] for each of the J sentences, for $\hat{\boldsymbol{\mu}}^*$, across the K individual signers. In other words, we take time-aligned sequences for a particular sentence and train the GP using the multiple productions of that sentence by the K signers. The trained model for a specific sequence j is denoted as $\boldsymbol{S}_j$:

$$\hat{\boldsymbol{\mu}}_t^{*,j} \sim \boldsymbol{S}_j := GP\left(mean^j(t), cov^j(t,t')\right), \tag{3}$$

where *mean* is a mean function and *cov* is a covariance function determining the covariance between any pair of timepoints t and t'. The GP therefore provides us with an approximation of the distribution of embedding trajectories for a particular sentence, across multiple signers.

To train the GP models, we utilise the negative of the marginal log likelihood (MLL) as our loss function. The negative MLL for the aligned latents $\hat{\boldsymbol{\mu}}^{*,j}$ given the inputs T_j is defined as:

$$
\begin{aligned}
\mathcal{L}_{\boldsymbol{GP}} &= -\log p(\hat{\boldsymbol{\mu}}^{*,j}|T_j^*) \\
&= -\log \mathcal{N}(\hat{\boldsymbol{\mu}}^{*,j}|mean^j, cov^j)
\end{aligned}
$$
$$
= -\frac{1}{2}\left(\hat{\boldsymbol{\mu}}^{*,j} - mean^j\right)^T \left(cov^j\right)^{-1}\left(\hat{\boldsymbol{\mu}}^{*,j} - mean^j\right) - \frac{1}{2}\log\left|cov^j\right| - \frac{N}{2}\log(2\pi)
$$
$$
(4)
$$

where $N = K_j T_j^*$, and K_j is the number of signers that produced sentence j. By taking the negative of the MLL, we maximise the log likelihood of the observed data under the GP model, thereby fitting the model to the data in a way that best explains the observed latents $\hat{\boldsymbol{\mu}}^{*,j}$.
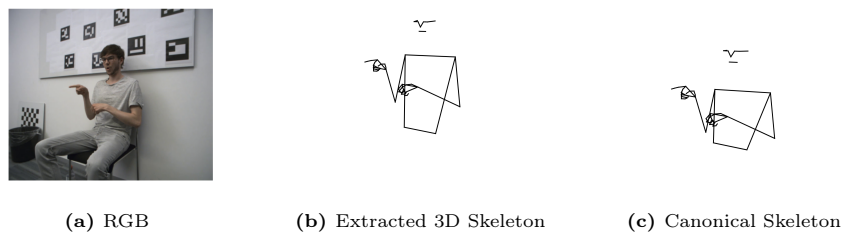
At inference time we take the embeddings for a test sequence for a specific j, $\hat{\boldsymbol{\mu}}_{test,j}$ and align it to the corresponding $\hat{\boldsymbol{\mu}}_{ref}$ such that it becomes $\hat{\boldsymbol{\mu}}^*_{test,j}$. We compare this sequence to the multivariate Gaussian posterior of the learnt model $\boldsymbol{S}_j$, returning principled uncertainty estimates for each t in the sequence.

4   Experiments

We evaluate our method using real-world SL data from native signers and language learners. We first outline our SL Sentence Repetition Test dataset and discuss the human rating scheme. We provide implementation details and compare our approach to the manual ratings by demonstrating quantitative and qualitative results.

4.1   Dataset

A recent study suggests that Sentence Repetition Tests (SRTs), which are widely used as a means of assessment for spoken language, can be applied to SL assessment [13]. SRTs ensure a comprehensive evaluation of signing ability; by requiring both comprehension and production, they provide a robust measure of language proficiency in the context of SL. During the testing process, each participant sees a prerecorded signed sequence video twice and is then asked to repeat it, i.e., the test taker has to comprehend, process, and produce language [57]. SRTs often work with a binary concept of correctness [57]. In this work a partial credit scale is used (as in [51]) in order to provide more informative feedback to the participant.

(a) RGB   (b) Extracted 3D Skeleton   (c) Canonical Skeleton

Fig. 2: Example frame from the dataset showing 2a the RGB frame of a participant from one of the camera views, 2b the uplifted 3D skeleton, and 2c the bone length adjusted canonical skeleton.

Table 1: Selection of examples from the Sentence Repetition Test

Sentence ID High German Written Sentence (English Translation)
A   Das Essen gestern Abend im Restaurant war schlecht.
(Last night the food in the restaurant was bad.)
E   Ich mag diesen Salat gar nicht.
(I don't like this salad at all.)
L   Er/Sie ist nicht da, weil er/sie krank ist.
(He/she is not there because he/she is sick.)

We create our Swiss-German Sign Language (Deutschschweizerische Gebärdensprache, DSGS) SRT dataset by recording a repetition test across 12 sentences of varying difficulty, determined by the number of signs, as well as morphological and syntactic complexity. The test is taken by a combination of 10 native signers and 14 language learners. Some examples of the sentences are shown in Tab. 1. We use the data from the native signers as our gold standard for training our model and the learners' data for evaluation.

We extract 3D canonical skeleton pose data (as shown in Fig. 2c) for the dataset, with each pose represented by 61 nodes in 3D Cartesian space. We sort the native signer data to include only sentences which are produced in the sentence order matching the initial reference and use this data for training the GPs model. We evaluate the model using all sentences produced by the language learners.

4.2   Manual Ratings

The data is analysed by eight native raters of DSGS using rating criteria designed to provide a comprehensive assessment of signing accuracy and fluency [17]. Raters are trained on a standardised rubric and evaluate videos of the sentences across six criteria: manual components, mouth components, eyebrow movements, head movements, eye gaze, and sentence structure.

Each criterion is assessed on a three-point scale, allowing for more nuanced feedback compared to a binary system. To ensure reliability, 14 videos are designated as anchor videos and are rated by all eight raters. The remaining 97 videos are assessed by two raters each in an overlapping design, with measures taken to balance video allocation and minimise potential bias. Analysis of the ratings provides inter-rater reliability, allowing us to determine the most reliable criterion for assessment.

For our experiments we choose to evaluate against the criterion for manual features on the sign produced by the language learners. Each rater provides a score for each manual component of the sign in the sentence. We take the mean of the ratings across the components in the sentence for each rater; and then take the mean across the raters that rated the sentence-learner pair. We repeat this for every sentence and learner. This provides a single score, 1 to 3, for each learner, for each sentence, that can be used for comparison with the output of our system.

4.3   Implementation Details

The encoder of our VAE consists of an input layer of size 183, followed by two hidden layers with sizes 100 and 50 perceptrons respectively. We implement fully connected layers and ReLU activation functions at the output of each layer except the final output layer, where we use a TanH function with its output scaled by a value of 6 to map the output of the network to the coordinate space of our pose data. The output of the encoder is split into two separate fully connected layers, each of size 10, representing the mean and log-variance of the latent space distribution. The mean and log-variance values are combined using the reparameterisation trick to calculate a 10-dimensional z-value vector. The decoder mirrors the structure of the encoder. It takes the 10-dimensional latent vector and passes it through two hidden layers, of size 50 and 100 respectively. The final output layer of the decoder reconstructs the original input dimension with size 183.

We initialize the weights of the fully connected layers using Kaiming normal initialization [15], with the biases initialized to 0.01. During training, we scale the added noise by a value of 0.001. For our loss function, Eq. (2), we choose an α of 0.9 and a β value of 0.0001 based on empirical experimentation. We train with a batch size of 32 for 100,000 epochs, with a learning rate of 0.001. We use Adam as our optimizer [25], and train over all canonical skeletons in the dataset.

For the DTW we choose a radius of size 20. For our GP Regression model we implement the 'ExactGP' model from GPyTorch [10]. We choose Gaussian likelihood as our likelihood function, use a Radial Basis Function as our kernel type, and initialise the mean function as a constant set to zero. We implement a gamma prior over the length scale with concentration and rate values both set to 0.1. We train with a learning rate of 0.1 until the loss reaches a threshold of 0.001.

4.4   Quantitative Results

In this section we evaluate the performance of our system using two distinct methods. The first method, which we refer to as the Probability Density method (PD method) is as follows. For each t in the test sequence, we calculate the probability density of the learner with respect to the learnt distribution at that point in the sequence for each latent dimension, resulting in a multidimensional mean. We then take the average of this mean, resulting in a single score for signing proficiency which we refer to as the Probability Density Measure (PD Measure). We expect a learner assigned a high manual rating to receive a high PD Measure and vice versa.

The second method quantifies the number of instances where the learner deviates from the distribution defined by the Motion Envelope, we refer to this as the Out of Distribution Count. Specifically, this method counts the occurrences where the learner falls outside the high confidence region, summing across all dimensions. The high confidence region is defined as the region that covers where we expect the true function values to lie with 95 percent probability [39]. This method is particularly effective at assessing anomaly detection. We expect a learner assigned a high manual rating to receive a low Out of Distribution Count and vice versa.

We standardise the resulting scores from our model and the manual ratings data using z-scoring standardisation. We apply the standardisation across each rater individually for all their ratings which increases the comparability of ratings from different raters. We then apply the standardisation on a per sentence level for the output scores of our system and the manual ratings. This results in standardised beta coefficients (which range from -1 to +1) when performing the regression analysis.

Linear Regression Analysis. We first evaluate our system by performing linear regression between the output scores of our model and the manual ratings data, measuring the standardised beta coefficient.

The results for the two methods can be seen in Tab. 2. A notable result here is the difference in scores when assessing using the PD Measure or Out of Distribution Events. For sentences A, B, C, E, G, I, K, L the first method achieves the best results, where as for sentences D, F, H, J the second method performs better. Both methods can be deemed useful. The PD method provides a more complete score over the entire sequence as all points in time are used in its calculation. However, it may be skewed negatively by acceptable deviations in sentence productions that are within distribution but far from the mean, as these will score relatively low compared to those with smaller distances to the mean of the distribution.

The Out of Distribution Count method only incorporates events into the score when the threshold is exceeded, providing a good method for anomaly detection, countering the downside of the PD method mentioned above.

For some of the sentences, the results for the PD measure and Out of Distribution Count are both low. One reason for this may be due to a non-linear

Table 2: Linear Regression Results. *Bolded* results for β, the *Standardised Beta Coefficient*, indicate the stronger correlation for each sentence out of the two methods. β represents the degree of correlation between the manual ratings and the outputs of the system.

Sentence	Prob. Density Measure ↑	Out of Dist. Count ↓
A	**0.60**	-0.15
B	**0.19**	0.03
C	**0.31**	-0.28
D	0.37	**-0.40**
E	**0.18**	-0.09
F	0.35	**-0.70**
G	**0.37**	-0.35
H	0.27	**-0.45**
I	**0.24**	-0.09
J	0.00	**-0.36**
K	**0.57**	-0.17
L	**0.46**	-0.45

Table 3: Spearman Rank Correlation Coefficient Results. *Bolded* results indicate the stronger correlation for each sentence out of the two methods.

Sentence	Prob. Density Measure ↑	Out of Dist. Count ↓
A	**0.69**	-0.19
B	**0.27**	-0.20
C	0.31	**-0.49**
D	0.35	**-0.42**
E	**0.30**	-0.15
F	0.43	**-0.60**
G	0.38	**-0.50**
H	0.22	**-0.51**
I	**0.20**	0.14
J	-0.03	**-0.43**
K	**0.56**	-0.26
L	**0.44**	-0.40

relationship between the manual ratings and the output of the system. To investigate this we present results using the Spearman Rank Correlation Coefficient.
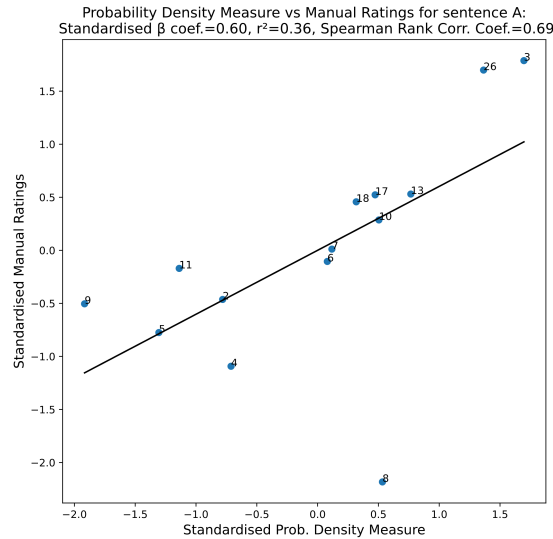
Spearman Rank Correlation Coefficient (SRCC). The SRCC is a measure of the strength and direction of the association between two variables that are assumed to be monotonic but not necessarily linear, based on the ranked values of the data.

In Tab. 3 we show that this metric offers complementary validity to that in Tab. 2 suggesting that the results are robust and not a spurious outcome of

metric choice. Furthermore, the strong SRCC scores shows that the monotonic relationship between the manual ratings and the system scores may be non-linear.

4.5   Plot Analysis

Fig. 3 showcases the model's strong agreement with the manual rating data for an example sentence. The model assigns low and high scores to the correct learners with respect to the manual ratings data, demonstrating its effectiveness in SL assessment. The correlation is strongly positive and almost linear.
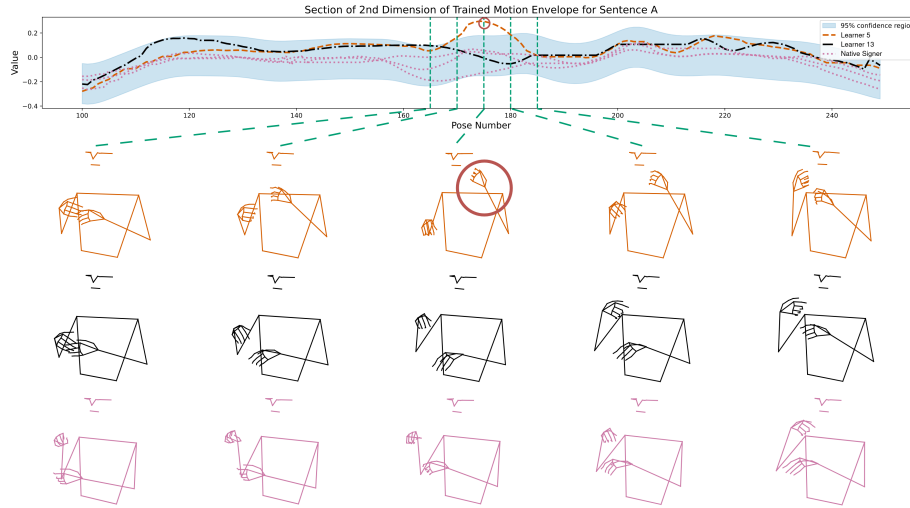

Fig. 3: Figure showing *standardised PD Measures* against the *standardised manual ratings* for *sentence A*. The *blue points* represent the language learners that produced the sentence, labelled with their predefined signer ID. The *black line* represents the line of best fit from the linear regression.

Learner 8 is a significant outlier in this plot, with our system assigning a mid-level score but being manually rated low. When looking at the Many Facets Rasch Measurement [37] for severity among raters, it becomes apparent that the sample is an outlier due to it being rated by the two most severe raters. In this case, the manual rating may be skewed negatively by their severity.

4.6   Qualitative Results

We now examine our system qualitatively, by using examples of a high and low scoring SL learner with respect to the learnt Motion Envelope and visualise their results.

Fig. 4: Top plot shows a section of from the latent dimension of the Motion Envelope *Confidence Region* with encoded SkeletonVAE signals overlayed for *Sentence A*. Below, decoded pose data for the latents is visualised for *Learner 5* (top), *Learner 13* (middle) and one *Native Signer* (bottom) for Pose Numbers 165-185 in steps of 5. The *red circle* indicates *Learner 5*'s peak deviation from the distribution.

As shown in Fig. 3, Learner 5 and Learner 13 both lie close to the line of best fit of the linear regression. Learner 5 receives a low overall single score from our system and is similarly rated by the manual raters whereas for Learner 13 the opposite is true, receiving high scores. As such these two language learners make a good example for further evaluation.

The plot on Fig. 4 shows time varying latent signals from one of the SkeletonVAE dimensions for Sentence A ranging from Pose Number 150 to 250 for learners and native signers against the learnt Motion Envelope confidence region. Learner 5 is shown leaving the learnt confidence region at pose number 170, with its greatest distance from the distribution occurring at 175 before returning to the distribution. On the contrary to this, Learner 13 stays within distribution throughout the sequence, coming close to the upper bounds at point but remaining within the confidence region, indicating its variation is acceptable.

This visualisation demonstrates the pipelines ability to temporally determine where anomalies have occurred, and by how far they differ from the learnt distribution over natural variations. The latent signals from three of the native signers used to train the Motion Envelope for this sentence are visualised to demonstrate examples of the natural variation in SL between deaf individuals.

The decoded pose sequences for the two learners and one of the native signers are displayed below the plot, focusing on the region where the anomaly occurs. We take the latent signals between Pose Numbers 165 and 185 and decode them

using the SkeletonVAE visualising every fifth pose within the range. At 165 the poses for the two learners are similar to each other, and slightly differ from the Native Signer, but stay within a margin of error. At 170, the arms of Learner 5 move in the opposite direction to Learner 13 and the native signer, who start to converge. At 175, Participant 13 is furthest in pose from the other two examples, with the wrong arm in the air. This is reflected as the point at which the participant is furthest from the learnt distribution. After this, the learner starts to move towards the direction of the learnt distribution, finally converging back with the other two examples as shown at pose number 185. This visualisation provides a spatial context of the error occurring in the skeleton space.

5   Conclusion

Sign Language Assessment tools are useful to aid in language learning and are underdeveloped. Previous work has focused on isolated signs, classification, or comparison against a single reference video to assess SL. In this paper, we proposed a novel assessment system to assess the comprehensibility of continuous SL sequences by modelling the natural distribution in human motion over multiple native deaf participants.

Our experiments demonstrated that modelling using multiple native signers can lead to robust and interpretable results. This approach can be used to provide visual feedback to users in spatio-temporal contexts to aid in SL learning and assessment. We evaluated our results using real data from language learners and showed strong correlation between manually rated data and our approach.

As future work, we would like to expand our system to include non-manual feature assessment as these are important linguistic features that modify the meaning of SL.

References

1. Arendsen, J., Lichtenauer, J.F., Holt, G.t., van Doorn, A.J., Hendriks, E.A.: Acceptability ratings by humans and automatic gesture recognition for variations in sign productions. In: 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition. pp. 1–6 (2008). https://doi.org/10.1109/AFGR.2008.4813347
2. Bansal, D., Ravi, P., So, M., Agrawal, P., Chadha, I., Murugappan, G., Duke, C.: Copycat: Using sign language recognition to help deaf children acquire language skills. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. CHI EA '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3411763.3451523

3. Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: A review for statisticians. Journal of the American Statistical Association **112**(518), 859–877 (Apr 2017). `https://doi.org/10.1080/01621459.2017.1285773`
4. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7784–7793 (2018). `https://doi.org/10.1109/CVPR.2018.00812`
5. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020). `https://doi.org/10.48550/arXiv.2003.13830`
6. Cui, R., Liu, H., Zhang, C.: Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1610–1618 (2017), `https://api.semanticscholar.org/CorpusID:7779968`
7. Duolingo: Duolingo - the world's best way to learn a language (2024), `https://www.duolingo.com`, accessed: 2024-06-13
8. Fang, S., Sui, C., Zhang, X., Tian, Y.: Signdiff: Learning diffusion models for american sign language production. arXiv preprint arXiv:2308.16082 (2023). `https://doi.org/10.48550/arXiv.2308.16082`
9. Feng, X., Lu, X., Si, X.: Taijiquan auxiliary training and scoring based on motion capture technology and dtw algorithm. International Journal of Ambient Computing and Intelligence **14**, 1–15 (01 2023). `https://doi.org/10.4018/IJACI.330539`
10. Gardner, J.R., Pleiss, G., Bindel, D., Weinberger, K.Q., Wilson, A.G.: Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In: Advances in Neural Information Processing Systems (2018). `https://doi.org/10.48550/arXiv.1809.11165`
11. Gong, J., Foo, L.G., He, Y., Rahmani, H., Liu, J.: Llms are good sign language translators. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18362–18372 (June 2024). `https://doi.org/10.48550/arXiv.2404.00925`
12. Grobel, K., Assan, M.: Isolated sign language recognition using hidden markov models. In: 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation. vol. 1, pp. 162–167 vol.1 (1997). `https://doi.org/10.1109/ICSMC.1997.625742`
13. Haug, T., Batty, A.O., Venetz, M., Notter, C., Girard-Groeber, S., Knoch, U., Audeoud, M.: Validity evidence for a sentence repetition test of swiss german sign language. Language Testing **37**(3), 412–434 (2020). `https://doi.org/10.1177/0265532219898382`
14. Haug, T., Mann, W., Knoch, U. (eds.): The Handbook of Language Assessment Across Modalities. Oxford University Press, Oxford (2022), `https://doi.org/10.1093/oso/9780190885052.001.0001`
15. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification (2015). `https://doi.org/10.48550/arXiv.1512.03385`
16. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-VAE: Learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations (2017), `https://openreview.net/forum?id=Sy2fzU9gl`
17. Holzknecht, F., Haug, T., Battisti, A., Tissi, K., Sidler-Miserez, S., Ebling, S.: Reforming sign language assessment: Setting up a longitudinal learner corpus of rated

elicited imitation performances to develop an ai-driven sign language assessment system (07 2024). `https://doi.org/10.13140/RG.2.2.32214.66886`

18. Holzknecht, F., Tornay, S., Battisti, A., Batty, A., Tissi, K., Haug, T., Ebling, S.: Automated sign language vocabulary assessment: Comparing human and machine ratings and studying learner perceptions. Language Assessment Quarterly pp. 1–21 (2024). `https://doi.org/10.1080/15434303.2024.2364877`

19. Ivashechkin, M., Mendez, O., Bowden, R.: Improving 3d pose estimation for sign language. In: 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). pp. 1–5 (2023). `https://doi.org/10.1109/ICASSPW59220.2023.10193629`

20. Jain, H., Harit, G.: An unsupervised sequence-to-sequence autoencoder based human action scoring model. In: 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP). pp. 1–5 (11 2019). `https://doi.org/10.1109/GlobalSIP45357.2019.8969424`

21. Jain, H., Harit, G., Sharma, A.: Action quality assessment using siamese network-based deep metric learning. IEEE Transactions on Circuits and Systems for Video Technology **31**(6), 2260–2273 (2021). `https://doi.org/10.1109/TCSVT.2020.3017727`

22. Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., Fu, Y.: Skeleton aware multi-modal sign language recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 3413–3423 (June 2021). `https://doi.org/10.48550/arXiv.2103.08833`

23. Kim, J.H., Hwang, E.J., Cho, S., Lee, D.H., Park, J.: Sign language production with avatar layering: A critical use case over rare words. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 1519–1528. European Language Resources Association, Marseille, France (Jun 2022), `https://aclanthology.org/2022.lrec-1.163`

24. Kindiroglu, A.A., Kara, O., Ozdemir, O., Akarun, L.: Transfer learning for cross-dataset isolated sign language recognition in under-resourced datasets. In: Proceedings of the 18th International Conference on Automatic Face and Gesture Recognition (FG 2024). Institute of Electrical and Electronics Engineers (IEEE) (2024). `https://doi.org/10.48550/arXiv.2403.14534`

25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017). `https://doi.org/10.48550/arXiv.1412.6980`

26. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2022). `https://doi.org/10.48550/arXiv.1312.6114`

27. Kingma, D.P., Salimans, T., Welling, M.: Variational dropout and the local reparameterization trick. In: Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015), `https://proceedings.neurips.cc/paper_files/paper/2015/file/bc7316929fe1545bf0b98d114ee3ecb8-Paper.pdf`

28. Koller, O., Bowden, R., Ney, H.: Automatic alignment of hamnosys subunits for continuous sign language recognition. In: LREC 2016: 10th edition of the Language Resources and Evaluation Conference. pp. 121 – 128 (2016)

29. Koller, O., Zargaran, O., Ney, H., Bowden, R.: Deep sign: Hybrid cnn-hmm for continuous sign language recognition. In: The British Machine Vision Conference (BMVC) (2016)

30. Koller, O., Ney, H., Bowden, R.: Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)

31. Koller, O., Zargaran, S., Ney, H.: Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017). `https://doi.org/10.1109/CVPR.2017.364`
32. Krebs, J., Malaia, E.A., Fessl, I., Wiesinger, H.P., Roehm, D., Wilbur, R., Schwameder, H.: Motion capture analysis of verb and adjective types in Austrian Sign Language (ÖGS). In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 11619–11624. ELRA and ICCL, Torino, Italia (May 2024), `https://aclanthology.org/2024.lrec-main.1015`
33. Kusters, A., De Meulder, M., O'Brien, D.: Innovations in deaf studies: Critically mapping the field. In: Innovations in deaf studies: The role of deaf scholars. vol. 12, pp. 1–53. Oxford University Press Oxford (2017)
34. Long-fei, C., Nakamura, Y., Kondo, K.: Modeling user behaviors in machine operation tasks for adaptive guidance (2020). `https://doi.org/10.48550/arXiv.2003.03025`
35. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., Chang, W.T., Hua, W., Georg, M., Grundmann, M.: Mediapipe: A framework for building perception pipelines (2019). `https://doi.org/10.48550/arXiv.1906.08172`
36. Morais, R., Le, V., Tran, T., Saha, B., Mansour, M., Venkatesh, S.: Learning regularity in skeleton trajectories for anomaly detection in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019). `https://doi.org/10.48550/arXiv.1903.03295`
37. Myford, C.M., Wolfe, E.W.: Detecting and measuring rater effects using many-facet rasch measurement: Part i. Journal of applied measurement **4**(4), 386–422 (2003)
38. Napier, J., Leeson, L.: Sign Language in Action. In: Sign Language in Action, pp. 50–84. Palgrave Macmillan UK, London (2016). `https://doi.org/10.1057/9781137309778_3`
39. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press (11 2005). `https://doi.org/10.7551/mitpress/3206.001.0001`
40. Salvador, S., Chan, P.: Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis **11**(5), 561–580 (2007). `https://doi.org/10.3233/IDA-2007-11508`
41. Sandler, W.: Prosody and syntax in sign languages. Transactions of the Philological Society **108**(3), 298–328 (2010). `https://doi.org/10.1111/j.1467-968X.2010.01242.x`
42. Sandler, W., Lillo-Martin, D.C.: Sign language and linguistic universals. Cambridge University Press (2006)
43. Saunders, B., Camgoz, N.C., Bowden, R.: Progressive transformers for end-to-end sign language production. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 687–705. Springer International Publishing (2020). `https://doi.org/10.1007/978-3-030-58621-8_40`
44. Selvaraj, P., Nc, G., Kumar, P., Khapra, M.: OpenHands: Making sign language recognition accessible with pose-based pretrained models across languages. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2114–2133. Association for Computational Linguistics, Dublin, Ireland (May 2022). `https://doi.org/10.18653/v1/2022.acl-long.150`

45. Starner, T., Forbes, S., So, M., Martin, D., Sridhar, R., Deshpande, G., Sepah, S., Shahryar, S., Bhardwaj, K., Kwok, T., Sehgal, D., Hassan, S., Neubauer, B., Vempala, S.A., Tan, A., Heath, J., Kumar, U.U., Mosur, P.V., Hall, T.M., Singh, R., Cui, C.Z., Cameron, G., Dane, S., Tanzer, G.: Popsign ASL v1.0: An isolated american sign language dataset collected via smartphones. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023), `https://openreview.net/forum?id=yEf8NSqTPu`
46. Stoll, S., Camgoz, N.C., Hadfield, S., Bowden, R.: Text2sign: towards sign language production using neural machine translation and generative adversarial networks. International Journal of Computer Vision **128**(4), 891–908 (2020). `https://doi.org/10.1007/s11263-019-01281-2`
47. Tamura, S., Kawasaki, S.: Recognition of sign language motion images. Pattern Recognition **21**(4), 343–353 (1988). `https://doi.org/10.1016/0031-3203(88)90048-9`
48. Tarigopula, N., Tornay, S., Muralidhar, S., Magimai Doss, M.: Towards accessible sign language assessment and learning. In: Proceedings of the 2022 International Conference on Multimodal Interaction. p. 626–631. ICMI '22, Association for Computing Machinery, New York, NY, USA (2022). `https://doi.org/10.1145/3536221.3556623`
49. Tornay, S., Bowden, R., Doss, M.M., Camgöz, N.C.: A phonology-based approach for isolated sign production assessment in sign language (20201025 - 20201029). `https://doi.org/10.1145/3395035.3425251`
50. Tornay, S., Nanchen, A., Battisti, A., Holzknecht, F., Tarigopula, N., Maldonado, O.M., Camgöz, N.C., Razavi, M., Tissi, K., Sidler-Miserez, S., Bream, P.B., Ebling, S., Haug, T., Bowden, R., Magimai-Doss, M.: Web smile demo: a web application providing automated feedback on sign language vocabulary production. In: 44th Language Testing and Research Colloquium: Language Assessment for a Global, Digital, and More Equitable Era. New York City, USA (June 7–9 2023)
51. Vinther, T.: Elicited imitation:a brief overview. International Journal of Applied Linguistics **12**(1), 54–73 (2002). `https://doi.org/10.1111/1473-4192.00024`
52. Walsh, H., Ravanshad, A., Rahmani, M., Bowden, R.: A data-driven representation for sign language production. In: Proceedings of the 18th International Conference on Automatic Face and Gesture Recognition (FG 2024). Institute of Electrical and Electronics Engineers (IEEE) (2024). `https://doi.org/10.48550/arXiv.2404.11499`
53. Walsh, H., Saunders, B., Bowden, R.: Changing the representation: Examining language representation for neural sign language production. In: Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives. pp. 117–124. European Language Resources Association, Marseille, France (Jun 2022). `https://doi.org/10.48550/arXiv.2210.06312`
54. Wen, H., Xu, Y.: Learning to score sign language with two-stage method (2024). `https://doi.org/10.48550/arXiv.2404.10383`
55. Wong, R., Camgoz, N.C., Bowden, R.: Sign2GPT: Leveraging large language models for gloss-free sign language translation. In: The Twelfth International Conference on Learning Representations (2024). `https://doi.org/10.48550/arXiv.2405.04164`
56. Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., Lu, J.: Finediving: A fine-grained dataset for procedure-aware action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2949–2958 (June 2022). `https://doi.org/10.48550/arXiv.2204.03646`

57. Yan, X., Maeda, Y., Lv, J., Ginther, A.: Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. Language Testing **33**(4), 497–528 (2016). https://doi.org/10.1177/0265532215594643
58. Yin, A., Zhong, T., Tang, L., Jin, W., Jin, T., Zhao, Z.: Gloss attention for gloss-free sign language translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2551–2562 (June 2023). https://doi.org/10.1109/CVPR52729.2023.00251
59. Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., Lei, Z., Zhang, D.: Gloss-free sign language translation: Improving from visual-language pretraining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20871–20881 (October 2023). https://doi.org/10.48550/arXiv.2307.14768