Co-Speech Gesture Detection through Multi-phase Sequence Labeling Supplementary Materials

Esam Ghaleb[1] Ilya Burenko[2,3] Marlou Rasenberg[4,6] Wim Pouw[5] Peter Uhrig[2,3]
Judith Holler[5,6] Ivan Toni[5] Aslı Özyürek[5,6] and Raquel Fernández[1]
[1]University of Amsterdam [2]ScaDS.AI Dresden/Leipzig [3]TU Dresden [4]Meertens Institute
[5]Radboud University [6]Max Planck Institute for Psycholinguistics
e.ghaleb@uva.nl raquel.fernandez@uva.nl

1. Dataset - Gesture Coding and Statistics

This section provides an overview of the gesture annotation process and the results of the inter-rater reliability check (for gesture identification and coding). The annotation was first leveraged by a study in Rasenberg *et al*. [5] to study how cross-speaker gestural and lexical alignment emerges.

Gesture Annotation In this dataset, only the stroke phase of co-speech gestures was annotated, which is the most meaning-bearing phase [2, 4]. Gestures were then divided into three categories: (1) iconic gestures that represent physical attributes or actions connected to a referent, (2) deictic gestures, which are often known as pointing gestures, and (3) other gestures, denoting predominantly beat and interactive gestures.

Statistics The fact that the dataset is collected in a referential game context made iconic gestures the most common category. In detail, the classification of annotated gestures is distributed as follows: a significant majority of 4952 iconic gestures, 360 of other types, and a relatively small count of 145 deictic gestures. Additionally, to distinguish between actual gestures and non-gestures, 642 movement segments were coded, encompassing self-adjustments or hand movements, as non-gestures. Regarding the length of the annotated strokes, the data indicates an average time of 0.58 seconds, with the most common duration being 0.24 seconds and a median value of 0.41 seconds.

Inter-rater Reliability The co-speech gesture coding procedures were conducted in two parts. The codings were assessed for inter-rater reliability based on trials from the total dataset and involved two independent coders. In the first part, 96 trials were coded, yielding 296 gesture annotations for comparison. The coders agreed on gesture identification

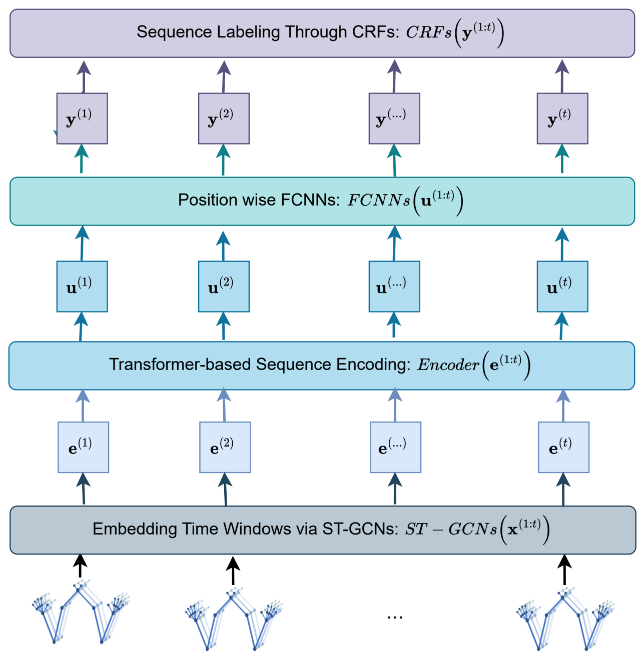

Figure 1. The architecture of the proposed model as outlined in Section 2.

89.2% of the time. A specialized Staccato algorithm [3] was used to account for variation in handedness, annotation length, and the number of segments, resulting in scores between 0.71 and 0.80 (on a scale from -1 to 1), suggesting similar segmentation understanding. The second part had a similar procedure, generating 406 gesture comparisons. The inter-rater agreement was slightly lower at 84.7%. The same Staccato algorithm was used to standardize segmentation, with the scores ranging between 0.61 and 0.77.

2. Implementation & Model Parameters

Implementation Details We train all our models with the same set of hyperparameters, which were identified through

Layer Type	Layer Details	Dimensions
ST-GCNs Model	Input ST-graph	$3 \times 27 \times 18$
	Model's 10 layers	Table 2
	Output	256
Transformer Encoder	*Positional encoding* layer input	256
	Positional encoding sequence max length	40
	Four stacked *Transformer Encoders* input	256
	Transformer Encoders Feed-forward networks	128×4
	Transformer Encoders Attention heads	8
	Output	256
FCNNs	Layer 1 Input	256
	Layer 1 Output	128
	Layer 3 Input	128
	Layer 3 Output	128
	Layer 5 Input	128
	Layer 5 Output	Number of labels

Table 1. Layer-wise structure and dimensions of the proposed model's components: ST-GCNs, Transformer Encoders, and the Fully Connected Neural Networks.

a process of random search. We use stochastic gradient descent with 0.1 base learning rate and an $L2$-regularization term with weight 10^{-4} to update models' weights. We increase the learning rate linearly for the first 20 epochs and divide it by 10 at the 50^{th} epoch. We train models for 80 epochs using 8 Nvidia A100 video cards with a batch size of 256, which was enough for the models to converge.

Model Layers and Dimensions The proposed model, as depicted in Figure 1, consists of the following components: (1) Spatio-Temporal Graph Convolutional Networks (ST-GCNs), (2) Transformer Encoders, (3) Position Wise Fully Connected Neural Networks (FCNNs), and (4) sequence labeler layer that employs Conditional Random Fields (CRFs). Table 1 lists the architecture's layers and their dimensions. Section 2 gives a brief overview of ST-GCNs, and Table 2 list our ST-GCNs layers and their dimensions. The full implementation of the model is available in the GitHub repository: `https://github.com/EsamGhaleb/Multi-Phase-Gesture-Detection`

ST-GCNs GCNs, a subset of graph neural networks, are models that have emerged from the success of traditional Convolutional Neural Networks (CNNs). They extend the convolution operation (template matching) from CNNs to accommodate data structured as graphs, allowing them to handle data with varying structures. GCNs are ideal for data structures that can be represented as graphs, such as spatio-temporal graphs of body joints, social networks or molecular structures [6]. In our research, GCNs prove particularly

Layer	In Channel	Out Channel	Stride
l1	3	64	None
l2	64	64	None
l3	64	64	None
l4	64	64	None
l5	64	128	2
l6	128	128	None
l7	128	128	None
l8	128	256	2
l9	256	256	None
l10	256	256	None

Table 2. ST-GCNs layers parameters and dimensions [1].

useful given the nature of our data, which uses ST-graphs to represent skeletal movements.

Technically, ST-GCNs extend conventional convolution operations to GNNs with features represented on a spatial graph V. The input feature map f_{in} at frame t is a c-dimensional vector for each node in the graph. For instance, in our study, c is a three-dimensional vector of each joint position (*i.e.* x and y) and the joint detection confidence at the input layer. The convolution operation is performed for each node v_i according to the formula: $f_{out}(v_i) = \sum_{v_j \in B_i} \frac{1}{Z_{ij}} f_{in}(v_j).w(l_i(v_j))$ where B_i represents the neighboring nodes v_j, Z_{ij} is a normalization term, w is a learnable kernel, and l_i maps the weight vectors for each vertex.

In the spatial context, ST-GCNs adopt spatial configuration partitioning to map weights, creating three subsets, denoted as K: the root node, a centripetal group, and centrifugal nodes. The convolution operation, in its vectorized form, is defined as $\boldsymbol{f}_{out} = \sum_k^{K_v} \boldsymbol{A}_k \odot \boldsymbol{M}k(\boldsymbol{f}in\boldsymbol{W}_k)$ where K_v is the kernel size, $\boldsymbol{A}_k$ is the adjacency matrix normalized by $\hat{D}^{-\frac{1}{2}}$, and $\boldsymbol{M}$ and $\boldsymbol{W}$ are learnable matrices. Temporal convolution is implemented with an $L \times 1$ convolutional layer to learn features from adjacent frames.

In our implementation, we used a pre-trained model for sign language recognition by Jiang *et al*. [1]. Table 2 lists the ST-GCNs model's layers and dimensions. Finally, a comprehensive description of this model can be found in the seminal work on ST-GCNs by Yan *et al*. [6].

References

[1] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3413–3423, 2021. 2, 3

[2] Adam Kendon. Gesture units, gesture phrases and speech. In *Gesture: Visible Action as Utterance*, chapter 7, pages 108–126. Cambridge University Press, 2004. 1

[3] Andy Lücking, Sebastian Ptock, and Kirsten Bergmann. Assessing agreement on segmentations by means of staccato, the segmentation agreement calculator according to thomann. In *International gesture workshop*, pages 129–138. Springer, 2011. 1

[4] David McNeill. *Hand and mind*. University of Chicago Press, 1992. 1

[5] Marlou Rasenberg, Asli Özyürek, Sara Bögels, and Mark Dingemanse. The primacy of multimodal alignment in converging on shared symbols for novel referents. *Discourse Processes*, 59(3):209–236, 2022. 1

[6] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 2, 3