# Geometry-aware Recognition of Mouth Articulations for Sign Language Understanding

Nurzhigit Ongalov ✉ and Bogdan Kwolek

AGH University of Krakow, 30 Mickiewicza, 30-059 Kraków, Poland
{ongalovna,bkw}@agh.edu.pl

**Abstract.** Recognition of vowel-related mouth articulations in continuous Japanese Sign Language (JSL) signing requires modeling fine-grained geometric structure and temporally coherent lip dynamics. We formulate this task as a structured spatio-temporal graph learning problem over lip landmark trajectories. A geometry-aware multi-branch ST-GCN architecture is proposed, operating on 41 nose-anchored facial landmarks and enriched with linear and angular motion descriptors to capture both spatial configuration and dynamic deformation. Experiments on a publicly available JSL dataset demonstrate that the proposed Lip-STGCN outperforms baseline ST-GCN and tree-based models under a strict cross-subject evaluation protocol, achieving a macro F1-score of approximately 63%. Ablation analysis confirms that jointly modeling structured geometry and motion dynamics is essential for robust vowel articulation recognition in continuous signing.

**Keywords:** Continuous Sign Language Recognition, Spatio-Temporal Graph Convolution, Geometry-Aware Modeling, Landmark-Based Motion Analysis

## 1 Introduction

Sign language is a visual language relying on coordinated manual and non-manual movements. Mouth actions constitute a linguistically essential component, contributing phonological, morphological, and semantic information that complements signed units [1]. In many sign languages, mouth articulations contribute linguistically relevant semantic and grammatical distinctions, thereby directly influencing interpretation [2]. Comprehensive surveys further emphasize the importance of modeling these non-manual components for reliable sign language recognition systems [3].

Continuous Sign Language (SL) recognition remains challenging, as lexical units merge without explicit boundaries, articulation varies across signers and contexts, and non-manual expressions interact tightly with manual components [4]. These challenges are amplified at the vowel level, where subtle spatio-temporal lip deformations must be distinguished under variability in speed, pose, and coarticulation, requiring representations that preserve both geometric structure and temporal continuity.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_37

Recent studies in facial expression and action unit recognition highlight the value of integrating geometric priors with adaptive weighting strategies. Xie et al. [5] showed that reweighting handcrafted feature losses improves discrimination by balancing heterogeneous geometric cues, while complementary graph-based formulations demonstrate that explicitly modeling relational dependencies among facial components enhances robustness and representation quality.

We propose a graph-based spatio-temporal framework for recognizing vowel-related mouth patterns in continuous signing videos. Landmark trajectories from *41 facial landmarks* (40 lip points and one nose anchor) are modeled using optimized ST-GCN architectures. The method is evaluated on the publicly available JSL dataset [1], refined to better satisfy ST-GCN requirements, comprising **6,133** labeled frames, with an ablation study assessing the impact of landmark subsets and motion descriptors. By jointly modeling geometric relations and temporal evolution, the framework enables vowel identification within continuous signing sequences.

The contributions are threefold: (i) formulation of vowel-level mouth articulation recognition in continuous signing as a structured spatio-temporal graph learning problem, with JSL used for empirical validation; (ii) a geometry-aware ST-GCN operating on nose-anchored lip landmarks to capture spatial structure and temporal dynamics; and (iii) systematic experimental evaluation on a publicly available JSL dataset with detailed ablation analysis.

## 2 Related Work

### 2.1 Geometry-Aware and Graph-Based Modeling

Visual speech recognition seeks to infer linguistic content from mouth motion, yet it is constrained by projecting high-dimensional articulatory dynamics onto low-dimensional, temporally compressed visual signals that are sensitive to pose and inter-speaker variability. Since facial motion follows coordinated anatomical constraints, it requires structured modeling rather than isolated appearance cues [6]. Recent graph-based approaches address this limitation. Zhao et al. [7] showed that embedding explicit geometric relations within attentive graph networks improves discrimination by modeling structured inter-landmark interactions. Wang et al. [8] demonstrated that geometry-aware graph convolutions enhance robustness through global contour encoding. Li et al. [9] introduced relation-guided graph representations to capture non-uniform semantic dependencies across facial regions, while Luo et al. [10] formalized heterogeneous edge interactions via multi-dimensional edge feature learning. Jiang et al. [11] show that adaptive graph construction and relation-aware connectivity improve attribute discrimination by modeling latent correlations between facial regions. At the architectural level, Guo et al. [12] provided evidence that hybrid structured–learnable graph formulations yield superior efficiency–performance trade-offs compared to unconstrained models.

Earlier visual speech systems relied on contour analysis and geometric deformation descriptors to represent mouth articulation as a linguistically structured

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_37

signal [13], yet these handcrafted approaches showed limited generalization. Quantitative evidence further supports structured temporal modeling: multi-scale temporal convolution improves word-level accuracy on LRW and LRW1000 from 84.1% to 85.3% and from 38.23% to 41.4%, respectively [6].

### 2.2 Temporal Modeling and Continuous Recognition

Temporal modeling imposes additional constraints: recurrent and dilated architectures enlarge receptive fields but may weaken short-range continuity critical for fine-grained vowel dynamics [12]. Because facial motion reflects coordinated temporal deformation, local continuity must be preserved alongside structured spatial relations [6].

Although geometry-aware graph models have advanced, they are typically validated on static or short sequences and do not explicitly model continuous articulatory transitions [7]. Relation-guided representations similarly emphasize short-term dependencies [9]. In Japanese Sign Language, mouth patterns are linguistically meaningful within continuous signing, where subtle geometric transitions encode semantic distinctions not recoverable from temporally sparse representations [1]. Similar conclusions have been reported for German Sign Language [2] and in broader reviews covering American, British, Chinese, and Indian Sign Languages [3]. Together, these findings underscore the need for a unified geometry-aware spatio-temporal framework that preserves fine-grained articulatory dynamics while maintaining long-term temporal stability.

## 3 Proposed Method

### 3.1 Problem Formulation

Formally, mouth pattern recognition in continuous signing is formulated as a structured spatio-temporal learning problem over dynamic landmark sequences. Let $\mathbf{X} \in \mathbb{R}^{T\times V\times d}$ denote a temporally ordered landmark segment of length $T$, where each time step $\mathbf{x}_t \in \mathbb{R}^{V\times d}$ represents the geometric features of $V$ facial landmarks with feature dimensionality $d$. The objective is to learn a mapping

$$f_\theta : \mathbb{R}^{T\times V\times d} \to \mathcal{C}, \tag{1}$$

such that

$$\mathbf{X} \in \mathbb{R}^{T\times V\times d}, \quad f_\theta(\mathbf{X}) = y, \tag{2}$$

where $T$ denotes the temporal segment length, $V$ the number of landmarks, $d$ the feature dimensionality, and $\mathcal{C}$ the set of vowel categories. The model operates on short temporally ordered landmark segments and produces a single class prediction $y \in \mathcal{C}$ for each segment. Although classification is performed at the segment level, the temporal structure within $X$ preserves articulatory continuity and coarticulatory dynamics across consecutive frames.

With landmark-based representations, each frame is a graph signal on a time-varying graph. Let $\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t)$, where $\mathcal{V} = \{v_1, \ldots, v_V\}$ are $V$ facial landmarks

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_37

(40 lip points and one nose anchor) and $E_t$ encodes semantic or spatio-temporal relations. The adjacency matrix $A_t \in \mathbb{R}^{V \times V}$ may be fixed, adaptive, or sample-dependent [9,12]. The observation becomes $\mathbf{X}_t \in \mathbb{R}^{V \times d}$, where $d$ is feature dimensionality.

Parameters $\theta$ are estimated by minimizing the segment-level classification loss

$$\mathcal{L}(\theta) = \ell\big(f_\theta(\mathbf{X}),\, y\big), \tag{3}$$

where $\ell(\cdot)$ denotes cross-entropy or focal loss. This objective jointly optimizes the spatial landmark representation, the structured relational graph modeling, and temporal feature aggregation within each segment of length $T$.

In Japanese Sign Language, subtle articulatory transitions carry linguistic meaning that cannot be recovered from temporally sparse or structurally unorganized representations [1]; however, existing geometry-aware and spatio-temporal models treat geometric structure, relational dynamics, and temporal continuity largely in isolation, motivating an integrated framework for continuous mouth pattern recognition.

### 3.2 Structured Multi-Branch Graph Architecture

The proposed framework operates on landmark-based motion representations extracted using MediaPipe facial landmarks [14]. After nose-anchored cropping, lip landmarks are expressed in a crop-local coordinate system with the nose tip as a fixed origin, reducing global translation variance and enforcing cross-subject geometric consistency.

Articulatory dynamics are modeled by temporally smoothing landmark trajectories and computing first-order linear velocities. Polar angles relative to the nose-centered frame and their temporal derivatives provide angular velocity components. Each landmark is thus represented by five features: spatial coordinates, linear velocities, and angular velocity, jointly encoding static geometry and dynamic deformation.

The resulting tensor is arranged in the standard ST-GCN format $[B, C, T, V]$, where $C = 5$ denotes feature channels, $T$ is the temporal length, and $V = 41$ corresponds to the number of facial landmarks used (40 lip points and one nose anchor). Temporal segments of 7, 9, and 11 frames are extracted using sliding windows, producing short articulatory contexts of approximately 0.3–0.5 seconds at 24–25 fps. This temporal granularity preserves local motion continuity while retaining sufficient contextual coverage for vowel discrimination. The overall architecture of the proposed model is illustrated in Figure 1.

*Structured Graph Formulation.* Unlike conventional ST-GCN models employing a single global adjacency matrix, the proposed architecture introduces a semantically structured multi-branch graph design. Lip landmarks define graph nodes $V$, and four distinct adjacency matrices encode complementary anatomical and functional relationships: (i) an upper outer lip loop, (ii) a lower outer lip loop, (iii) an inner lip self-connectivity structure, and (iv) cross-region connections linking corners, nose, and chin anchors.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_37

Formally, each branch operates on a fixed topology $G^{(k)} = (V, E^{(k)})$ with adjacency matrix $A^{(k)} \in \mathbb{R}^{V \times V}$, where $k \in \{\text{upper}, \text{lower}, \text{inner}, \text{cross}\}$. This decomposition enables localized relational modeling of distinct articulatory regions while preserving structural priors.

Each branch consists of stacked spatial–temporal graph convolution blocks. Spatial propagation is performed via

$$\mathbf{X}' = \mathbf{X}A^{(k)}, \tag{4}$$

followed by $1 \times 1$ channel mixing and a temporal convolution with kernel size 7 to model motion evolution across frames. Residual connections ensure stable feature refinement. After two stacked blocks per branch, adaptive global pooling aggregates features into compact regional descriptors.

*Branch Fusion and Classification.* The outputs of the four branches are concatenated and fed into a fully connected fusion layer for classification. This structure promotes region-specific feature learning before global integration, assuming that upper lip, lower lip, inner lip, and cross-anchor interactions contribute differently to vowel articulation.

For comparison, non-graph baselines such as Random Forest and XGBoost are evaluated on aggregated sequence descriptors to assess the benefit of structured spatio-temporal modeling over non-relational approaches.

Collectively, the architecture reflects the premise that continuous vowel recognition relies on structured regional geometry and temporally coherent deformation dynamics, with semantically partitioned graph propagation providing a principled integration mechanism.

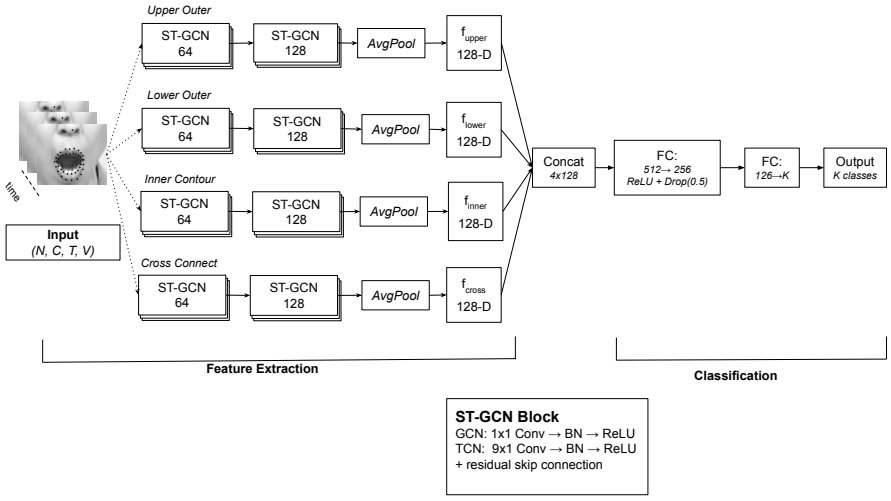


**Fig. 1.** Proposed Model Architecture.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_37

## 4 Handcrafted Lip Landmark Features

Let $\mathbf{X} \in \mathbb{R}^{B\times T\times V\times 2}$ denote a batch of landmark sequences, where $B$ is the batch size, $T$ is the number of frames in a temporal window, $V = 41$ is the number of tracked facial landmarks, and each landmark is represented by its 2-D screen coordinates $(x, y)$.

We write the landmark tensor for a single sample as

$$\mathbf{X}_b \in \mathbb{R}^{T\times V\times 2}, \quad b = 1, \ldots, B. \tag{5}$$

The $v$-th landmark at frame $t$ is denoted $\mathbf{p}_{t,v} = (x_{t,v}, y_{t,v}) \in \mathbb{R}^2$. Table 1 lists the semantically meaningful landmark indices used throughout this section.

**Table 1.** Landmark index mapping (0-based internal index → MediaPipe ID).

| Internal Index | Anatomical Location | MediaPipe ID |
|---|---|---|
| 0 | Nose tip (reference anchor) | 1 |
| 1 | Left mouth corner | 61 |
| 11 | Right mouth corner | 291 |
| 36 | Upper lip center | 13 |
| 26 | Lower lip center | 14 |
| 6 | Chin tip | 17 |

### 4.1 Rigid Alignment (Procrustes)

Raw landmark coordinates are subject to inter-frame head motion. To isolate lip-intrinsic deformation we apply a rigid (translation + rotation) alignment that maps every frame onto the coordinate system of the first frame $t = 0$.

### 4.2 Translation Removal

For each frame $t \in \{0, \ldots, T-1\}$ and the reference frame $t = 0$, we subtract the nose-tip coordinate (index 0) to center both frames:

$$\tilde{\mathbf{p}}_{t,v} = \mathbf{p}_{t,v} - \mathbf{p}_{t,0}, \quad v = 0, \ldots, V-1. \tag{6}$$

stacking landmarks into matrices $\tilde{\mathbf{P}}_0, \tilde{\mathbf{P}}_t \in \mathbb{R}^{V\times 2}$, the centered coordinates form the input to the rotation step.

**Rotation Removal via Orthogonal Procrustes** We seek the orthogonal matrix $\mathbf{R}_t \in \mathbb{R}^{2\times 2}$ ($\mathbf{R}_t^\top \mathbf{R}_t = \mathbf{I}$, $\det(\mathbf{R}_t) = +1$) that best aligns the current frame to the reference:

$$\mathbf{R}_t = \underset{\mathbf{R}:\, \mathbf{R}^\top\mathbf{R}=\mathbf{I}}{\arg\min} \left\|\tilde{\mathbf{P}}_t\mathbf{R} - \tilde{\mathbf{P}}_0\right\|_F^2. \tag{7}$$

The closed-form solution follows from the Singular Value Decomposition (SVD) of the cross-covariance matrix:

$$\mathbf{H}_t = \tilde{\mathbf{P}}_t^\top \tilde{\mathbf{P}}_0 \in \mathbb{R}^{2\times 2}, \qquad \mathbf{H}_t = \mathbf{U}\,\boldsymbol{\Sigma}\,\mathbf{W}^\top. \tag{8}$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_37

The optimal rotation is

$$\mathbf{R}_t = \mathbf{W}\mathbf{U}^\top. \tag{9}$$

The aligned landmark matrix for frame $t$ is then

$$\hat{\mathbf{P}}_t = \tilde{\mathbf{P}}_t\,\mathbf{R}_t \in \mathbb{R}^{V\times 2}. \tag{10}$$

After alignment, the full tensor $\hat{\mathbf{X}} \in \mathbb{R}^{B\times T\times V\times 2}$ contains pose-normalised coordinates with translation and in-plane rotation removed.

### 4.3 Static Geometric Features

The following scalar features are computed per frame $t$ from $\hat{\mathbf{P}}_t$ using a fixed set of landmark pairs. Let us introduce shorthand for the semantically important points:

$$\mathbf{l}_t = \hat{\mathbf{p}}_{t,1}, \quad \mathbf{r}_t = \hat{\mathbf{p}}_{t,11}, \quad \mathbf{u}_t = \hat{\mathbf{p}}_{t,36}, \quad \mathbf{d}_t = \hat{\mathbf{p}}_{t,26}, \quad \mathbf{o}_t = \hat{\mathbf{p}}_{t,0}, \quad \mathbf{c}_t = \hat{\mathbf{p}}_{t,6}. \tag{11}$$

**Normalisation Reference: Mouth Width** A scale-invariant reference is established from the inter-corner distance:

$$\mathrm{mw}(t) = \|\mathbf{l}_t - \mathbf{r}_t\|_2 + \varepsilon, \quad \varepsilon = 10^{-6}. \tag{12}$$

All features derived from distances in the vertical or depth directions are divided by $\mathrm{mw}(t)$ to remove subject-level scaling.

**Feature Set A — The Eight Geometric Scalars** For each frame $t$ we extract the following eight features $\mathbf{g}_t = [g_t^{(1)}, \ldots, g_t^{(8)}]^\top \in \mathbb{R}^8$.

**F1. Mouth Width**

$$g_t^{(1)} = \|\mathbf{l}_t - \mathbf{r}_t\|_2. \tag{13}$$

Absolute inter-corner horizontal span (unnormalised). Used both as a raw feature and as the denominator $\mathrm{mw}(t)$.

**F2. Normalised Mouth Openness**

$$g_t^{(2)} = \frac{\|\mathbf{u}_t - \mathbf{d}_t\|_2}{\mathrm{mw}(t)}. \tag{14}$$

Vertical lip gap relative to mouth width; a compact measure of jaw aperture and lip parting.

**F3. Normalised Jaw Openness**

$$g_t^{(3)} = \frac{\|\mathbf{o}_t - \mathbf{c}_t\|_2}{\mathrm{mw}(t)}. \tag{15}$$

Nose-to-chin Euclidean distance divided by mouth width. Captures gross mandibular depression independently of lip parting.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_37

**F4. Vertical Asymmetry**

$$g_t^{(4)} = l_{t,y} - r_{t,y}, \tag{16}$$

where $l_{t,y}$ and $r_{t,y}$ are the $y$-coordinates of the left and right corners respectively. A non-zero value indicates a tilted or asymmetrically raised lip corner.

**F5. Upper Lip Curvature**

$$g_t^{(5)} = \|\mathbf{l}_t + \mathbf{r}_t - 2\,\mathbf{u}_t\|_2. \tag{17}$$

This is the $\ell_2$ norm of the vector from the midpoint of the two corners to the upper-center landmark, doubled. Geometrically it equals twice the deviation of the upper-center from the chord connecting the two corners. A large value denotes a pronounced Cupid's bow.

**F6. Lower Lip Curvature**

$$g_t^{(6)} = \|\mathbf{l}_t + \mathbf{r}_t - 2\,\mathbf{d}_t\|_2. \tag{18}$$

Analogous to F5 for the lower lip.

**F7. Left Corner Angle**

$$g_t^{(7)} = \angle\Big(\overrightarrow{\mathbf{l}_t\mathbf{u}_t},\, \overrightarrow{\mathbf{l}_t\mathbf{d}_t}\Big) = \operatorname{atan2}((\mathbf{v}_1 \times \mathbf{v}_2),\ (\mathbf{v}_1 \cdot \mathbf{v}_2))\,, \tag{19}$$

where $\mathbf{v}_1 = \mathbf{u}_t - \mathbf{l}_t$, $\mathbf{v}_2 = \mathbf{d}_t - \mathbf{l}_t$, and the 2-D cross product is $v_{1x}v_{2y} - v_{1y}v_{2x}$. This captures the interior angle at the left corner of the mouth.

**F8. Right Corner Angle**

$$g_t^{(8)} = \operatorname{atan2}((\mathbf{v}_{1r} \times \mathbf{v}_{2r}),\ (\mathbf{v}_{1r} \cdot \mathbf{v}_{2r}))\,, \tag{20}$$

where $\mathbf{v}_{1r} = \mathbf{u}_t - \mathbf{r}_t$, $\mathbf{v}_{2r} = \mathbf{d}_t - \mathbf{r}_t$. Analogous to F7 for the right corner.

### 4.4 Temporal Dynamics Features

Given the per-frame geometric feature sequence $\{\mathbf{g}_t\}_{t=0}^{T-1}$ with $\mathbf{g}_t \in \mathbb{R}^8$, we enrich with first- and second-order finite differences to capture motion and acceleration of articulation.

**Velocity (First Difference)**

$$\dot{\mathbf{g}}_t = \mathbf{g}_t - \mathbf{g}_{t-1}, \quad t = 1, \ldots, T-1. \tag{21}$$

For $t = 0$ we set $\dot{\mathbf{g}}_0 = \mathbf{0}$ (zero-padding).

**Acceleration (Second Difference)**

$$\ddot{\mathbf{g}}_t = \dot{\mathbf{g}}_t - \dot{\mathbf{g}}_{t-1}, \quad t = 2, \ldots, T-1. \tag{22}$$

For $t \in \{0, 1\}$ we set $\ddot{\mathbf{g}}_t = \mathbf{0}$.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_37

**Concatenated Temporal Feature Vector** The full per-frame descriptor is the concatenation of position, velocity, and acceleration:

$$\mathbf{f}_t = \left[\mathbf{g}_t^\top,\ \dot{\mathbf{g}}_t^\top,\ \ddot{\mathbf{g}}_t^\top\right]^\top \in \mathbb{R}^{24}, \quad t = 0, \ldots, T-1. \tag{23}$$

### 4.5 Temporal Pooling

To produce a fixed-length representation from the variable-length sequence $\{\mathbf{f}_t\}_{t=0}^{T-1}$, we apply first- and second-moment temporal pooling:

$$\boldsymbol{\mu} = \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{f}_t \in \mathbb{R}^{24}, \qquad \boldsymbol{\sigma} = \sqrt{\frac{1}{T}\sum_{t=0}^{T-1}(\mathbf{f}_t - \boldsymbol{\mu})^2} \in \mathbb{R}^{24}. \tag{24}$$

The final feature vector is:

$$\boxed{\phi(\mathbf{X}) = \left[\boldsymbol{\mu}^\top,\ \boldsymbol{\sigma}^\top\right]^\top \in \mathbb{R}^{48}.} \tag{25}$$

## 5 Experimental Results

This section presents the experimental evaluation of the proposed framework. Evaluation metrics are defined in Subsection 5.1, followed by the dataset description in Subsection 5.2 and the training configuration in Subsection 5.3.

Baseline results obtained with a Random Forest classifier are reported first, followed by the performance of the proposed ST-GCN model. An ablation study then quantifies the contribution of architectural and feature components, and additional visual analyses examine feature structure and model decision behavior.

### 5.1 Evaluation Metrics

Performance is evaluated using accuracy, precision, recall, and F1-score. Due to pronounced class imbalance, macro F1-score is used as the primary model selection criterion, as it assigns equal weight to each class and prevents dominance by the majority non-vowel category. Accuracy is reported for completeness, while precision and recall provide additional class-wise insight.

Experiments follow the subject-disjoint split defined in Subsection 5.2. Reproducibility is ensured by fixing the random seed (42) across NumPy, Python, PyTorch (CPU and CUDA), and DataLoader workers, with deterministic CUDA enabled.

Temporal samples are constructed using sliding windows of length $T = 7$ and stride 3, producing overlapping segments while preserving local temporal coherence. Median labeling is applied unless stated otherwise, reducing sensitivity to short transitional noise. Pure non-vowel sequence dropping is disabled.

Training employs AdamW (learning rate 0.001, weight decay 0.0001) with a ReduceLROnPlateau scheduler (factor 0.5, patience 2) monitoring validation

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_37

loss. Early stopping (patience 7) is based on validation macro F1, with a maximum of 30 epochs and batch size 16. Class imbalance is addressed through `WeightedRandomSampler` with inverse-frequency weights and Focal Loss ($\gamma = 2$) with class weighting to emphasize hard examples. Overall, the protocol enables fair cross-subject evaluation and controlled comparison of structured graph-based configurations for continuous vowel articulation recognition.

### 5.2 Dataset

The dataset is derived from the publicly available JSL corpus introduced in [1], with a refined lip landmark subset to enhance geometric consistency and alignment stability while preserving the original annotation scheme. The preprocessing pipeline was further updated to incorporate nose-anchored normalization and consistent frame-level crop alignment, improving cross-subject spatial coherence. After preprocessing and filtering, the dataset contains **6,133** labeled frames from eight recordings featuring ten participants. Each frame is manually annotated into one of **six classes**: a non-vowel category and five vowel classes (*A, I, U, E, O*). The class distribution remains imbalanced, with the non-vowel category forming the majority due to the inherent structure of continuous signing.

For evaluation, subject-disjoint splits are employed to ensure cross-subject generalization. Specifically, subjects 1–5 are used for training, subject 6 for validation, and subjects 7–10 for testing. This protocol prevents identity leakage between training and testing sets.

### 5.3 Experimental Evaluation

This subsection reports the quantitative evaluation of the proposed framework on the cross-subject test split, analyzing overall performance, class-wise behavior, and error distribution patterns across configurations.

The confusion matrix of the best-performing graph-based spatio-temporal model is presented in Figure 2. The non-vowel class achieves the highest true positive count (209), followed by *I* and *A* with 94 and 63 correct predictions, respectively, while *U*, *E*, and *O* show lower true positives and stronger mutual confusions, particularly between *U*–*O* and *A*–*I*, indicating partial geometric–temporal overlap. These errors arise from strong geometric similarity between certain vowel configurations, with differences mainly reflected in subtle variations of lip opening and curvature. Short temporal segments and coarticulation further blur class boundaries, making closely related articulations difficult to distinguish.

Although 3D convolutional models can yield higher absolute performance in RGB-based settings, they rely on full facial imagery and retain identifiable visual information. In contrast, the proposed landmark-based graph representation uses only geometric keypoints, inherently reducing identity exposure and supporting privacy-preserving modeling, which is especially important for sign language datasets involving publicly recognizable individuals.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_37

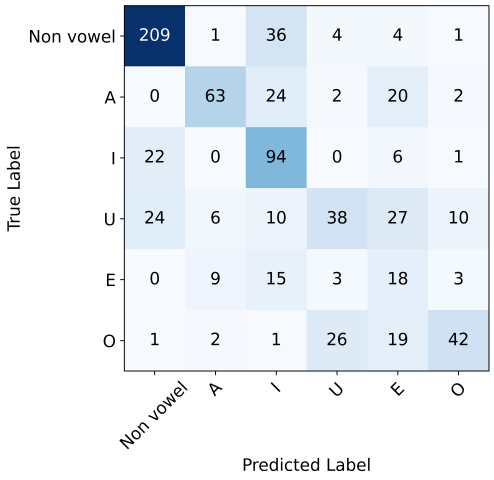


**Fig. 2.** Confusion matrix on predictions of the Lip-STGCN. Darker blue cells indicate higher prediction counts, highlighting the strongest classification performance along the main diagonal.

**Baseline Results: Random Forest** Figure 3 shows the relative importance of the top 20 geometric–temporal features derived from the Random Forest model. Static geometric descriptors clearly dominate, with mean mouth width, mean jaw opening, and mean mouth opening ranking highest, followed by angular measures and lower lip curvature, indicating that global mouth configuration is the primary discriminative cue. Mid-ranked features comprise dispersion statistics of mouth opening and angular values together with several first-order velocity descriptors, confirming that temporal variability complements static structure. Acceleration-based features appear at the lowest ranks, reflecting a comparatively smaller contribution of second-order dynamics relative to static geometry and linear motion.

**Graph-Based Results: ST-GCN** The comparative performance of all models is summarized in Table 2. The baseline ST-GCN reported in [1], which was trained using a five-point landmark subset, yields the lowest result with an F1-score of 49.52%, indicating limited capability in modeling fine-grained vowel dynamics. Tree-based ensembles improve upon this baseline: XGBoost achieves a mean F1-score of 52.88% over 20 runs, while Random Forest reaches 56.83% F1-score and attains notably high precision (68.67%), though with moderate recall.

The proposed Lip-STGCN achieves the strongest performance, with an F1-score of approximately 63%, an accuracy of about 62%, and a precision of roughly 66%, substantially surpassing the baseline ST-GCN. These findings demonstrate a consistent progression from conventional ST-GCN to tree-based methods and ultimately to the geometry-aware multi-branch Lip-STGCN, confirming the ad-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_37

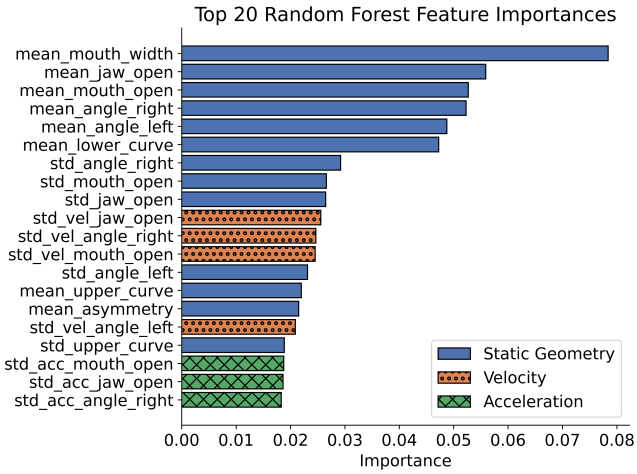


**Fig. 3.** Top 20 feature importance scores estimated by the Random Forest classifier for the extracted lip descriptors. Static geometric features (solid), velocity features (dotted), and acceleration features (cross-hatched) are differentiated by pattern. The Random Forest model used 500 estimators with *class_weight="balanced"*; remaining parameters followed scikit-learn defaults.

vantage of structured regional graph modeling with enriched dynamic features for continuous vowel articulation recognition.

**Table 2.** Performance comparison of classification models on the dataset for dynamic vowel recognition in Japanese Sign Language. Results are reported in terms of accuracy, precision, recall, and F1-score. For XGBoost and Random Forest, values denote mean performance over 20 independent runs.

| | Accuracy [%] | Precision [%] | Recall [%] | F1-score [%] |
|---|---|---|---|---|
| ST-GCN [1] | 51.91 | 55.27 | 51.91 | 49.52 |
| XGBoost | $55.37 \pm 2.1$ | $59.18 \pm 2.6$ | $55.37 \pm 2.1$ | $52.88 \pm 2.0$ |
| RF | $55.68 \pm 2.0$ | $68.67 \pm 2.4$ | $55.68 \pm 2.0$ | $56.83 \pm 1.9$ |
| **Lip-STGCN** | **62.45** | **66.30** | **62.45** | **62.90** |

**Ablation Study** An ablation study evaluated the impact of filter depth and feature composition on Lip-STGCN. The weakest results occur with the smallest configuration (32, 16) using only spatial coordinates $(x, y)$ or linear velocities $(x, y, v_x, v_y)$, both reaching an F1-score of approximately 51%, indicating that limited capacity and restricted features fail to capture discriminative articulatory dynamics (Table 3).

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_37

Using only spatial features in the larger (128, 64) configuration yields 55.36% F1, confirming that static geometry alone is insufficient. Adding angular velocity without linear motion provides only modest gains, yielding 54.39% F1 for the (128, 64) configuration with $(x, y, ang\_vel)$, whereas combining spatial and linear motion improves performance, reaching 60.22% F1 for the (64, 32) configuration with $(x, y, v_x, v_y)$, respectively. The highest performance is obtained with the full feature set $(x, y, v_x, v_y, ang\_vel)$ under (128, 64), reaching 62.90% F1, 62.45% accuracy, and 66.30% precision; notably, even the compact (32, 16) configuration achieves 62.03% F1 when all motion features are included. This suggests that feature richness can compensate for reduced model capacity.

These findings show that spatial features alone are inadequate, linear velocities substantially enhance discrimination, and integrating linear and angular motion provides the most consistent improvements, emphasizing the importance of jointly modeling geometric structure and dynamic articulation.

**Table 3.** Ablation study of Lip-STGCN under different filter configurations and feature sets.

| Filters | Features | Acc. [%] | Pre. [%] | Rec. [%] | F1 [%] |
|---|---|---|---|---|---|
| Base 512, 256 | $x, y$ | 61.18 | 65.27 | 61.04 | 61.12 |
| | $x, y, v_x, v_y$ | 55.82 | 62.74 | 56.03 | 55.91 |
| | $x, y, v_x, v_y, ang\_vel$ | 58.37 | 60.41 | 58.12 | 58.79 |
| | $x, y, ang\_vel$ | 58.21 | 61.36 | 58.07 | 58.94 |
| 128, 64 | $x, y$ | 56.14 | 55.48 | 56.02 | 55.36 |
| | $x, y, v_x, v_y$ | 61.07 | 61.42 | 61.18 | 59.26 |
| | $x, y, v_x, v_y, ang\_vel$ | **62.45** | **66.30** | **62.45** | **62.90** |
| | $x, y, ang\_vel$ | 57.26 | 64.08 | 57.11 | 54.39 |
| 64, 32 | $x, y$ | 60.83 | 65.12 | 60.94 | 59.48 |
| | $x, y, v_x, v_y$ | 60.18 | 66.74 | 60.03 | 60.22 |
| | $x, y, v_x, v_y, ang\_vel$ | 59.36 | 63.05 | 59.14 | 57.28 |
| | $x, y, ang\_vel$ | 60.27 | 64.33 | 60.18 | 58.41 |
| 32, 16 | $x, y$ | 54.82 | 58.16 | 55.03 | 51.28 |
| | $x, y, v_x, v_y$ | 55.11 | 57.42 | 55.08 | 51.37 |
| | $x, y, v_x, v_y, ang\_vel$ | **62.08** | **63.27** | **62.14** | **62.03** |
| | $x, y, ang\_vel$ | 57.19 | 58.34 | 57.05 | 53.12 |

To analyze the structure of the handcrafted geometric–temporal feature space, a two-dimensional t-SNE projection of the extracted lip descriptors is shown in Figure 4, separately for the training set ($n = 1418$) and the test set ($n = 813$). The embeddings exhibit partial clustering of vowel categories, with some classes forming compact regions and others showing overlap across both splits. The non-vowel class appears more dispersed, reflecting its heterogeneous articulatory behavior. Notably, a similar global structure is preserved between training and test distributions, indicating that the descriptors generalize consistently across unseen subjects. Although complete separation is not achieved, the observed grouping suggests that the descriptors encode meaningful class-dependent variation in articulatory dynamics while maintaining cross-subject stability.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_37

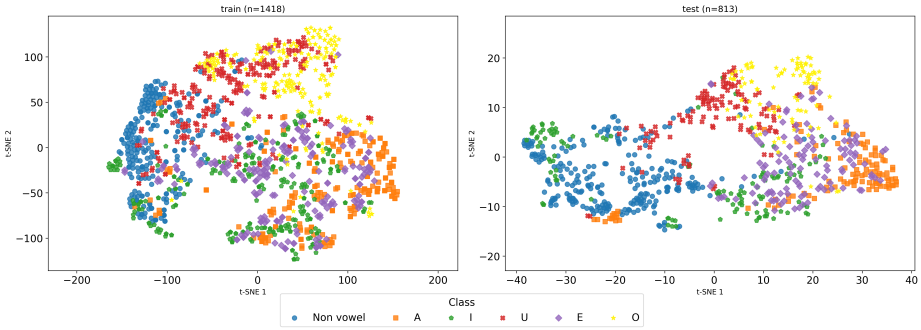


**Fig. 4.** t-SNE projection of the extracted lip feature vectors for the training ($n = 1418$) and test ($n = 813$) sets. The embedding illustrates the distribution and relative separability of the six vowel classes in the handcrafted geometric-temporal feature space. The figure is best interpreted in color to facilitate clear differentiation between class markers.

The system has been implemented in Python using PyTorch [15]. The neural networks were trained using CUDA on a Linux PC equipped with two Nvidia A100 GPUs. Using a batch size of $B = 16$, inference on the test set ($n = 813$) required 1.74 seconds in total, including feature extraction, corresponding to 2.14 ms per sequence.

## 6 Conclusions

In this work, we presented a geometry-aware spatio-temporal framework for recognition of vowel-related mouth articulations in continuous JSL signing. Vowel identification in continuous signing scenarios was formulated as a structured graph learning problem over nose-anchored lip landmark trajectories and modeled using a multi-branch ST-GCN architecture. The proposed Lip-STGCN was evaluated on a publicly available JSL dataset using a strict subject-disjoint cross-subject protocol. Experimental results demonstrated consistent improvements over baseline ST-GCN and tree-based ensemble models, with ablation analysis confirming the importance of integrating geometric structure and motion dynamics. The findings indicate that structured regional graph modeling provides an effective and interpretable solution for continuous vowel articulation recognition in JSL.

**Acknowledgment.** This work was supported by the National Science Center, Poland (NCN), under grant no. 2024/55/B/ST6/01580. We gratefully acknowledge Poland's high-performance computing infrastructure PLGrid (ACC Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018381.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_37

## References


1. Umeda, Y., Ongalov, N., Sroka, G., Shinji, S., Kwolek, B.: Continuous recognition of mouth patterns in Japanese Sign Language for visual communication. In: ACIIDS, Springer (2025) 115–128
2. Von Agris, U., Knorr, M., Kraiss, K.F.: The significance of facial features for automatic sign language recognition. In: Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), IEEE (2008) 1–6
3. Banerjee, K., Vats, I., Akhtar, N., KG, H., Kumar, A., Kumar, P., Dasila, P., Gautam, P., et al.: A review on artificial intelligence based sign language recognition techniques. In: Proc. 5th Int. Conf. Contemp. Comput. Informatics (IC3I), IEEE (2022) 2195–2201
4. Brock, H., Farag, I., Nakadai, K.: Recognition of non-manual content in continuous Japanese Sign Language. Sensors **20**(19) (2020) 5621
5. Xie, W., Shen, L., Duan, J.: Adaptive weighting of handcrafted feature losses for facial expression recognition. IEEE Trans. Cybern. **51**(5) (2019) 2787–2800
6. Martinez, B., Valstar, M.F., Jiang, B., Pantic, M.: Automatic analysis of facial actions: A survey. IEEE Trans. Affect. Comp. **10**(3) (2017) 325–347
7. Zhao, R., Liu, T., Huang, Z., Lun, D.P., Lam, K.M.: Geometry-aware facial expression recognition via attentive graph convolutional networks. IEEE Trans. Affect. Comp. **14**(2) (2021) 1159–1174
8. Wang, S., Zhao, A., Lai, C., Zhang, Q., Li, D., Gao, Y., Dong, L., Wang, X.: GCANet: Geometry cues-aware facial expression recognition based on graph convolutional networks. J. King Saud Univ. Comput. Inf. Sci. **35**(7) (2023) 101605
9. Li, G., Zhu, X., Zeng, Y., Wang, Q., Lin, L.: Semantic relationships guided representation learning for facial action unit recognition. In: Proc. AAAI. Volume 33. (2019) 8594–8601
10. Luo, C., Song, S., Xie, W., Shen, L., Gunes, H.: Learning multi-dimensional edge feature-based AU relation graph for facial action unit recognition. In: IJCAI. (2022) 1239–1246
11. Jiang, F., Huang, Q., Mei, X., Guan, Q., Tu, Y., Luo, W., Huang, C.: Face2Nodes: Learning facial expression representations with relation-aware dynamic graph convolution networks. Inf. Sci. **649** (2023) 119640
12. Guo, X., Polania, L., Zhu, B., Boncelet, C., Barner, K.: Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases. In: Proc. IEEE/CVF WACV. (2020) 2921–2930
13. Antonakos, E., Roussos, A., Zafeiriou, S.: A survey on mouth modeling and analysis for sign language recognition. In: Proc. IEEE FG. Volume 1., IEEE (2015) 1–7
14. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C., Yong, M.G., Lee, J., Chang, W., Hua, W., Georg, M., Grundmann, M.: MediaPipe: A framework for building perception pipelines. CoRR **abs/1906.08172** (2019)
15. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, L., Lin, Z., Antiga, A., Lerer, A.: PyTorch: An imperative style, high-performance deep learning library. In: Adv. Neural Inf. Process. Syst. (NeurIPS). Volume 32. (2019)



ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_37