A Resource and Evaluation Method for Phonological Continuity in Japanese Sign Language

Jundai Inoue, Daisuke Hara, Makoto Miwa
Toyota Technological Institute
2-12-1 Hisakata, Tempaku-ku, Nagoya, Aichi, 468-8511, Japan
{sd24410, daisuke, makoto-miwa}@toyota-ti.ac.jp

Abstract

Computational models for sign language processing often represent phonological components as categories. This approach, however, does not adequately capture the continuous nature of sign articulation, obscuring nuanced phonetic variation. Furthermore, the field has lacked resources and standardized methods to evaluate a model's ability to represent this continuity. In this work, we address these limitations. First, we introduce the JSL Ordered Triplet Dataset, a new manually-annotated resource designed to benchmark the modeling of gradual phonological progressions in Japanese Sign Language. Second, we propose a learning framework that reframes the task from classification to ranking, using Positive-Unlabeled (PU) learning to optimize the Area Under the ROC Curve (AUC). Our intrinsic evaluation on the new dataset shows that the learned continuous embeddings significantly outperform a cross-entropy baseline in ordering intermediate forms, improving the average accuracy on the continuity ranking task across phonological components from 81.52% to 91.71%. These embeddings also maintain strong discriminative power for standard component classification. This work provides the community with a valuable resource and a method for learning and evaluating more linguistically-grounded representations of sign language.

Keywords: Sign Language Processing, Language Resources, Evaluation, Phonology, Representation Learning

1. Introduction

The field of Sign Language Processing (SLP) has seen significant advancements, particularly in applications such as sign language recognition and translation (Zuo et al., 2023; Gong et al., 2024). Similar to spoken languages, sign languages are complex linguistic systems where meaning is constructed through the systematic combination of smaller, non-meaningful phonological components such as handshape, location, and movement (Stokoe, 1960). The ability to represent these components in a linguistically grounded manner is a foundational prerequisite for many downstream SLP tasks (Tavella et al., 2022; Kezar et al., 2023).

However, a representational gap exists between current deep learning-based SLP models and the continuous nature of sign language production in practical use. While sign language production is a continuous and analog signal, nearly all existing models treat its constituent parts as mutually exclusive, categories (Tavella et al., 2022; Kezar et al., 2023). This gap is illustrated in Figure 1. The standard classification-based methods, shown in the top panel, impose boundaries on continuous phenomena by assigning them to fixed categories. This process can lead to a loss of nuanced information about intermediate forms, phonetic variation, and phonotactic constraints. This loss of detail may limit the performance and linguistic fidelity of downstream applications, particularly for tasks that rely on subtle phonotactic rules.

To address this representational limitation, we depart from the standard classification approach. Instead, we propose to learn a continuous phonological embedding space that preserves the relational structure between components, as depicted in the bottom panel of Figure 1. The central research question we address is how a continuous representation of phonology, which reflects the graded nature of articulation, can be learned from datasets that provide only categorical labels. To this end, we first introduce the JSL Ordered Triplet Dataset, a new language resource specifically designed to assess a model's ability to represent the graded progression between phonological forms. By providing triplets of signs (v_1, v_2, v_3) where v_2 is an intermediate form between anchors v_1 and v_3, this resource enables a new, more nuanced form of intrinsic evaluation.

Given this new evaluation framework, we reformulate the task from a classification problem to a ranking problem by introducing a learning framework based on Positive-Unlabeled (PU) AUC optimization for this task. We hypothesize that by ranking all instances of one category higher than all instances[1] of others, this objective will encourage the model to learn a smooth manifold capturing the continuous structure of the phonological space.

Our contributions are summarized as follows:

- We introduce the JSL Ordered Triplet Dataset, a new resource for evaluating phonological continuity in sign language.

[1] Each instance corresponds to a sign video.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9514–9524
11–16 May 2026.

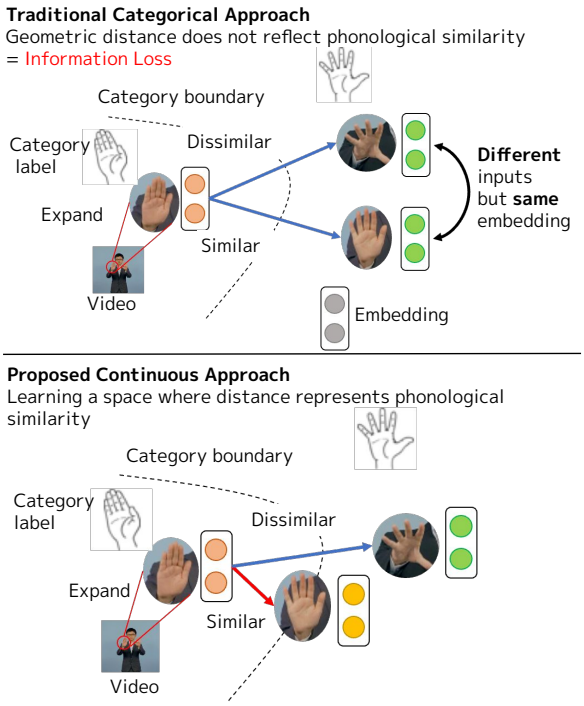


Figure 1: Conceptual comparison between categorical and continuous representations of sign language phonology. **Top**: Traditional categorical approach assigns phonological components to categories with hard boundaries, which cannot represent intermediate forms and gradual transitions. **Bottom**: Our proposed continuous approach learns a smooth manifold where phonologically similar forms are positioned closer together, preserving information about intermediate variations and enabling more linguistically-grounded representations. Icons from McKee et al. (2011).

- We propose a novel intrinsic evaluation metric, Continuity Ranking Accuracy, enabled by this dataset.
- We demonstrate that a PU-AUC optimization framework, evaluated on our new resource, learns superior continuous representations compared to a standard classification baseline, while maintaining competitive performance on traditional classification tasks.

2. Related Work

2.1. Sign Language Phonology

Sign languages are fully-fledged natural languages, yet they still differ from spoken languages in that they lack a universally adopted written form. Early linguistic analysis of sign languages was hindered by the absence of a systematic notation. Stokoe (1960) provided the first rigorous phonological analysis by decomposing signs into three components—location, handshape, and movement. This notation allowed signs to be described in a dictionary and laid the foundation for sign language phonology. In particular, subsequent systems such as the Hamburg Notation System (HamNoSys) and SignWriting extend this idea, offering more iconic glyphs and broader cross-linguistic applicability (Sutton, 1990; Hanke, 2004). While these notations treat phonological components as categories, linguistic research has long noted that parameters like handshape, location, and movement form a continuum in physical space (Goldin-Meadow and Brentari, 2017). Our work builds on this observation by preserving the continuous nature of articulation in a learned embedding space, rather than assigning signs to categories.

2.2. Computational Modeling of Sign Language Phonology

Deep learning approaches to sign language processing increasingly incorporate phonological information. For example, Kezar et al. (2023) improve isolated sign language recognition by using handshape, location, and movement as auxiliary prediction tasks within a multitask training setup. Tavella et al. (2022) introduce WLASL-LEX, a dataset annotated with multiple phonological properties, to encourage phonology-aware modeling. While these efforts demonstrate the value of phonological information, they still operate within a categorical framework: phonological features serve as labels for classification. Our work departs from this approach by treating phonology as a continuous manifold and learning representations that reflect graded similarity relationships. This allows us to capture subtle phonetic variations and transitions that are not represented in categorical approaches.

2.3. Sign Language Corpora and Resources

Research in Sign Language Processing (SLP) has been significantly advanced by the development of large-scale language resources. These can be categorized into lexical databases, designed for analyzing the abstract structure of the language, and corpora, which capture natural language use.

A landmark lexical database is ASL-LEX (Sehyr et al., 2021), which contains detailed phonological and psycholinguistic annotations for thousands of American Sign Language (ASL) signs. To enable a more precise analysis of the abstract properties of each sign and to control for variability, all signs were recorded by a single native signer. This controlled dataset has been instrumental in quantitative research on the structure of the ASL lexicon.

More recently, hybrid datasets that combine linguistic features with large-scale video data have emerged to facilitate machine learning research. A prominent example is WLASL-LEX (Tavella et al., 2022), which was created by aligning the phonological information from ASL-LEX with the WLASL video corpus, a collection of sign language videos featuring over 100 different signers. This resource was specifically designed to train models for the automatic recognition of phonological properties from video, leveraging the natural variations present in a multi-signer corpus.

While these existing resources provide invaluable phonological labels, they lack the fine-grained annotations of continuous progression necessary for evaluating the geometric properties of embedding spaces. Our work directly addresses this resource gap by providing a targeted dataset for this purpose.

2.4. Ranking from Positive and Unlabeled Data

Positive-Unlabeled (PU) learning addresses scenarios where only positive examples are labeled, while the remaining unlabeled data is a mix of positive and negative instances (Bekker and Davis, 2020). A primary challenge is that naively treating all unlabeled instances as negative introduces bias, as the unlabeled set contains hidden positive instances. Sakai et al. (2018) proposed an unbiased estimator for AUC optimization from PU data by decomposing the true Positive-Negative risk into observable Positive-Unlabeled and Positive-Positive components. While this framework has been applied to ranking and classification tasks, we adapt it to learn continuous phonological embeddings for sign language.

3. The JSL Ordered Triplet Dataset

To evaluate whether the learned embedding space captures continuous structure, we constructed a new evaluation dataset, the JSL Ordered Triplet Dataset. This dataset is designed to assess the model's ability to represent phonological continuity. This section details the dataset's construction process and statistics.

3.1. Dataset Construction

The dataset was constructed by two trained master's students as part of their research activities. The overall process involved sourcing data from the JSL Syllable Database (Yawata et al., 2017), creating comparison sets, and annotating ordinal relationships based on a detailed scheme.

Data Sourcing and Set Creation The construction process focused on creating small, comparable sets of videos for each phonological component (e.g., handshape). The procedure was as follows:

1. Two distinct anchor classes were selected (e.g., for handshape, a fully open hand vs. a closed fist).
2. Three representative videos were chosen from each of the two classes, creating a *sample set* of six videos.
3. This process was repeated to build a large number of sample sets for each component.

Each video clip in the dataset averages about 300 frames in length, including those selected for triplets.

Annotation Guidelines and Triplet Creation An overview of the annotation system is shown in Figure 2. From each six-video sample set, annotators were tasked with identifying a triplet of videos (v_1, v_2, v_3) that exhibited a clear, continuous phonological progression from the first anchor class towards the second. If no such clear progression could be found, the set was skipped. The judgment of "continuous progression" was guided by specific criteria for each component:

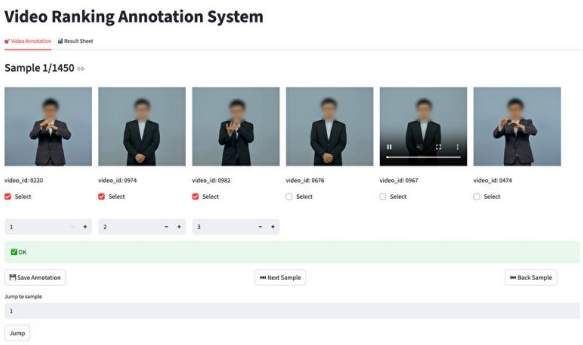


Figure 2: Overview of the annotation system.

Handshape Continuity was evaluated based on three axes of articulation:

- **Degree of Hand Opening/Closing:** The gradual change from a fully extended hand to a closed fist (Figure 3a).
- **Degree of Finger Flexion:** The degree to which fingers bend, for instance at the second knuckle.
- **Degree of Finger Spread:** The gradual change from fingers spread wide apart to being tightly closed together.

Component	Train	Dev
HS (Pre)	341	50
HS (Post)	632	39
Location	201	43
Movement	212	25
Total	1,385	157

Table 1: Annotation statistics for the JSL Ordered Triplet Dataset. Each triplet targets a single phonological component. HS: Handshape; Pre/Post: Pre-change/Post-change.

Location Continuity was defined by spatial changes along two axes:

- **Depth:** Movement towards or away from the signer's body, such as from a non-contact location to one touching a body part (Figure 3b).
- **Horizontal/Vertical Location:** Gradual changes in hand location within the signing space.

Movement Continuity was defined by gradual changes in the *direction of movement*, for example, a progression from a horizontal rightward movement to an upward-right diagonal movement (Figure 3c).

3.2. Dataset Statistics and Inter-Annotator Agreement

We follow the predefined train/dev split of the JSL Syllable Database (Yawata et al., 2017) and construct the corresponding splits of the triplet dataset by forming triplets only from videos within each split (Table 1). Note that the split names (Train/Dev) solely reflect the source partition of the JSL Syllable Database, and the triplet dataset is used for evaluation only, not for model training. To assess the reliability of the annotations, we measured inter-annotator agreement on a randomly sampled subset of 152 triplets from the train set. Two trained annotators independently ordered the same triplets, and their rankings were compared. The agreement metrics are shown in Table 2. The high agreement scores (84.21% triplet order agreement and 0.88 Pearson correlation) indicate that the continuous phonological progressions can be reliably identified by trained annotators.

3.3. Resource Availability

The dataset will be made available upon publication to encourage further research in this area.

Metric	Score
Triplet Order Agreement	84.21%
Pearson Correlation	0.88

Table 2: Inter-annotator agreement on 152 triplets from the training set.

4. Methodology

Our goal is to learn an embedding space that captures the continuous similarity relationships among the phonological components of sign language. However, a challenge exists: we lack direct continuous supervision for phonological similarity, as traditional datasets only provide categorical labels. To address this challenge, we depart from the standard approach of *classifying* components into categories and propose a *ranking*-based approach that models phonological similarity as a geometric relationship in a continuous space. We achieve this by using a Positive-Unlabeled (PU) learning framework to optimize the Area Under the ROC Curve (AUC) with a pairwise ranking loss, as illustrated in Figure 4. While this does not directly optimize for continuous similarity, the ranking objective encourages phonologically similar sign videos to be positioned closer in the embedding space. This encourages the formation of a continuous manifold structure suitable for linguistic representation.

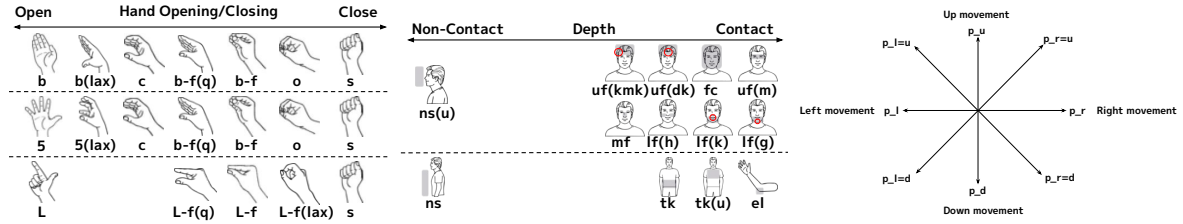
4.1. Model Architecture

We aim to learn a mapping function $f_\theta : \mathcal{X} \to \mathbb{R}^d$ from sign language videos to a continuous embedding space where phonological similarity is reflected by geometric proximity. To implement this mapping, the function f_θ is instantiated using a Video Vision Transformer (ViT-Base/16) (Wang et al., 2023) encoder combined with an MLP projection head. The ViT encoder extracts spatio-temporal features, which are then projected into the d-dimensional continuous embedding space.

4.2. Learning Framework with PU-AUC Optimization

To construct the desired continuous embedding space, we propose a learning framework based on Positive-Unlabeled (PU) AUC optimization. By training the model to consistently rank instances of the same phonological category higher than instances of different categories across the entire dataset, we hypothesize that intermediate instances will be positioned between clusters, thus encouraging the desired continuity.

We frame our learning objective using a multi-class extension of PU learning. Given a set of



(a) Handshape (b) Location (c) Movement

Figure 3: Representative examples of phonological continuity in hand articulation. (a) Handshape: Gradual change in hand opening. (b) Location: Gradual change in the horizontal position of the hand. (c) Movement: Gradual change in the direction of hand movement. Other axes of continuity described in the text (e.g., finger flexion, horizontal/vertical) are not shown due to space constraints.

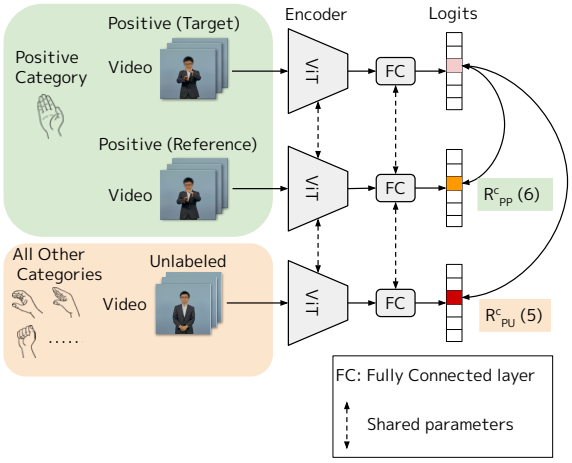


Figure 4: Overview of the proposed representation learning framework. An input sign video is processed by a ViT encoder, followed by a fully-connected (FC) projection head which outputs per-class logits. The PU-AUC loss is computed using the logits corresponding to the positive category, optimizing the model to rank positive instances higher than unlabeled instances. This ranking objective encourages the formation of a continuous phonological manifold where the geometric distance between embeddings reflects their similarity. Note: This diagram illustrates the process for Handshape. The same process is applied independently to Location and Movement components.

phonological categories $\mathcal{C}$, we adopt a One-vs-Rest approach. For each class $c \in \mathcal{C}$, we treat samples belonging to that class as "Positive" (P) and all other samples as "Unlabeled" (U). The objective is to maximize the AUC for each class. The ideal loss function for this task would be the true PN risk, which minimizes the ranking loss between true positive ($x_P \sim D_c$) and true negative ($x_N \sim D_{\neg c}$) examples:

$$\mathcal{R}^c_{\mathsf{PN}}(g) = \mathbb{E}_{x_P \sim D_c, x_N \sim D_{\neg c}}[l(g_c(x_P) - g_c(x_N))] \quad (1)$$

where l is a pairwise ranking loss function and $g_c(x)$ is the logit for class c. In our implementation, we use the logistic loss: $l(z) = \log(1 + e^{-z})$. However, since we only have access to unlabeled data U, we cannot directly minimize Eq. Equation (1). Following the theory of PU learning, this unknown PN risk can be estimated unbiasedly from observable P and U data (Sakai et al., 2018). The distribution of unlabeled data D_U is a mixture of positive and negative distributions: $D_U = \pi_c D_c + (1 - \pi_c) D_{\neg c}$, where π_c is the class prior of c. Based on this, the expected risk between P and U samples, $\mathcal{R}^c_{\mathsf{PU}}(g)$, can be decomposed as:

$$\mathcal{R}^c_{\mathsf{PU}}(g) = \pi_c \mathcal{R}^c_{\mathsf{PP}}(g) + (1 - \pi_c)\mathcal{R}^c_{\mathsf{PN}}(g) \quad (2)$$

where $\mathcal{R}^c_{\mathsf{PP}}(g) = \mathbb{E}_{x_P, x_{P'} \sim D_c}[l(g_c(x_P) - g_c(x_{P'}))]$ is the positive-positive risk. Rearranging Equation (2) to isolate $R^c_{PN}(g)$, we obtain an unbiased estimator of the PN risk expressed solely in terms of observable P and U quantities. We refer to this estimator as $R^c_{\mathsf{PU\text{-}AUC}}$:

$$\mathcal{R}^c_{\mathsf{PU\text{-}AUC}}(g) = \frac{\mathcal{R}^c_{\mathsf{PU}}(g)}{1 - \pi_c} - \frac{\pi_c \mathcal{R}^c_{\mathsf{PP}}(g)}{1 - \pi_c} \quad (3)$$

The first term represents the observable ranking loss between P and U pairs, while the second term corrects for the bias introduced by the presence of positive instances within the unlabeled set. In practice, we approximate this expected risk with an empirical risk $\hat{\mathcal{R}}(g)$ computed over a mini-batch B. Let C_B be the set of unique classes in B, B_c be the set of samples of class c in B (with size n_c), and $B_{\neg c}$ be the set of other samples in B. The total empirical risk is:

$$\hat{\mathcal{R}}(g) = \sum_{c \in C_B} \left(\frac{\hat{\mathcal{R}}^c_{\mathsf{PU}}}{1 - \pi_c} - \frac{\pi_c \hat{\mathcal{R}}^c_{\mathsf{PP}}}{1 - \pi_c} \right) \quad (4)$$

where the component terms are calculated as:

$$\hat{\mathcal{R}}^c_{\mathsf{PU}} = \frac{1}{n_c|B_{\neg c}|} \sum_{x_i \in B_c, x_j \in B_{\neg c}} l(g_c(x_i) - g_c(x_j)) \tag{5}$$

$$\hat{\mathcal{R}}^c_{\mathsf{PP}} = \frac{1}{n_c(n_c-1)} \sum_{x_i, x_j \in B_c, i \neq j} l(g_c(x_i) - g_c(x_j)) \tag{6}$$

This ranking objective encourages the model to learn a smooth phonological manifold rather than category boundaries.

5. Evaluation

We designed an evaluation protocol to validate our proposed method. The experiments aim to answer two key questions: (1) To what extent does the learned embedding space capture human perceptual similarity (intrinsic evaluation)? and (2) Do these continuous embeddings maintain discriminative power on a standard classification task?

5.1. Datasets

- **JSL Syllable Database:** For our primary training and evaluation, we use the JSL Syllable Database (Yawata et al., 2017). This dataset contains 1,078 videos, each treated as an instance, performed by a single native JSL signer and annotated with phonological components (handshape, location, movement). We use this dataset to train the model and to evaluate phonological component classification performance. We split the data into 755 instances for training, 161 for development, and 162 for testing. Given the single-signer nature of this dataset, our results reflect the model's ability to capture phonological distinctions within one consistent signing style, which may limit generalizability across signers.
- **JSL Ordered Triplet Dataset:** For our intrinsic evaluation, we use our new JSL Ordered Triplet Dataset introduced in §3 (Inoue et al., 2026), which contains 1,385 triplets in the train split and 157 in the dev split. We report evaluation results on both splits.

5.2. Models and Training

The goal is to learn a phonological embedding space. We train the ViT encoder in a single-task setup to predict only the phonological components of a sign. We compare two models, differing only in their loss function:

- **ViT-CE (Baseline):** The model is trained to classify phonological components using a standard categorical Cross-Entropy (CE) loss. This represents the conventional approach.
- **ViT-PU-AUC (Ours):** The model is trained using our proposed PU-AUC ranking loss, as described in §4.2.

We use the AdamW Schedule-Free optimizer (Defazio et al., 2024) with a learning rate of 1.0×10^{-4} and apply layer decay to the ViT encoder. Full hyperparameter settings are provided in §A. All experiments were conducted using NVIDIA RTX 6000 Ada Generation, NVIDIA RTX PRO 6000 Blackwell Max-Q, and NVIDIA GeForce RTX 4060 Ti GPUs. Our ViT encoder is initialized with the `vit_b_k710_dl_from_giant.pth` weights, which were released as part of the VideoMAE v2 project[2].

Class Prior Setting Rather than estimating class priors π_c from the small training dataset (which could introduce noise), we treat them as hyperparameters. Based on the theoretical framework of PU-AUC optimization, we set uniform priors $\pi_c = 1/|\mathcal{C}|$ for all classes c, where $|\mathcal{C}|$ is the number of phonological categories. This approach avoids potential estimation errors while maintaining the theoretical soundness of the PU-AUC framework.

5.3. Evaluation Protocol

Our evaluation is two-fold. We conduct an intrinsic evaluation to assess the quality of the learned representations themselves, and a classification task to measure their discriminative power.

5.3.1. Intrinsic Evaluation: Continuity Ranking Accuracy

To evaluate the quality of the learned phonological representations, we measure their ability to model continuous phonological progressions. This novel metric directly assesses whether the learned embedding space correctly orders intermediate instances along a linguistically meaningful axis. For a given triplet (v_1, v_2, v_3) that is annotated as exhibiting a continuous phonological progression:

1. Extract the embedding vectors e_1, e_2, e_3.
2. Define a "continuity axis" vector, $V_{\text{axis}} = e_3 - e_1$, using the endpoints e_1 and e_3.
3. To evaluate if the intermediate point e_2 is projected between e_1 and e_3 on this axis, we first transform e_2 into a relative vector from the origin e_1, denoted as $e'_2 = e_2 - e_1$.

[2] https://github.com/OpenGVLab/VideoMAEv2

4. We then project e_2' onto V_{axis} and compute a score as the ratio of the projected vector's length:

$$\text{Score} = \frac{e_2' \cdot V_{\text{axis}}}{\|V_{\text{axis}}\|^2} \quad (7)$$

5. A triplet is considered correctly ordered if the score satisfies $0 < \text{Score} < 1$.

6. The final evaluation metric is the accuracy over all triplets in the train and dev sets.

5.3.2. Phonological Component Classification

To assess whether the learned representations maintain discriminative power, we evaluate their performance on the standard task of phonological component classification. We report classification accuracy on the development set.

6. Results and Discussion

Our experiments are designed to answer two key questions. First, does our proposed PU-AUC optimization framework learn a continuous representation of phonology from labels? (Intrinsic Evaluation). Second, do these continuous embeddings maintain discriminative power for standard classification?

6.1. Intrinsic Evaluation: Capturing Phonological Continuity

To validate our central hypothesis, we first evaluate whether the learned embedding space correctly orders intermediate phonological forms along a continuous axis. We use our newly created JSL Ordered Triplet Dataset and measure the Continuity Ranking Accuracy, which assesses if the model captures the graded progression between phonological anchors. This is the primary evaluation of our proposed method.

The results, presented in Table 3, support the effectiveness of our approach. Our ViT-PU-AUC model consistently outperforms the ViT-CE classification baseline. For instance, the PU-AUC model achieves 98.24% on the train set and 94.00% on the dev set for pre-change handshape compared to the baseline's 78.30% and 84.00%, respectively. While the performance on the dev set is more varied, the PU-AUC model generally maintains an advantage. This suggests that the ranking objective of PU-AUC optimization encourages the formation of continuous structure. While the classification baseline can learn to separate the anchor classes, it does not consistently order the intermediate instances that lie between them. In contrast, our method learns a smooth manifold that reflects the continuous nature of phonological transitions. The consistent performance between train and dev sets for our method suggests that this learned continuity generalizes to unseen data.

Component	Method	Train	Dev
HS (Pre)	CE	78.30	84.00
	PU-AUC	**98.24**	**94.00**
HS (Post)	CE	77.53	82.05
	PU-AUC	**88.77**	**97.44**
Location	CE	63.18	76.74
	PU-AUC	**85.07**	**81.40**
Movement	CE	74.88	84.00
	PU-AUC	**95.73**	**96.00**
Average	CE	75.23	81.52
	PU-AUC	**91.62**	**91.71**

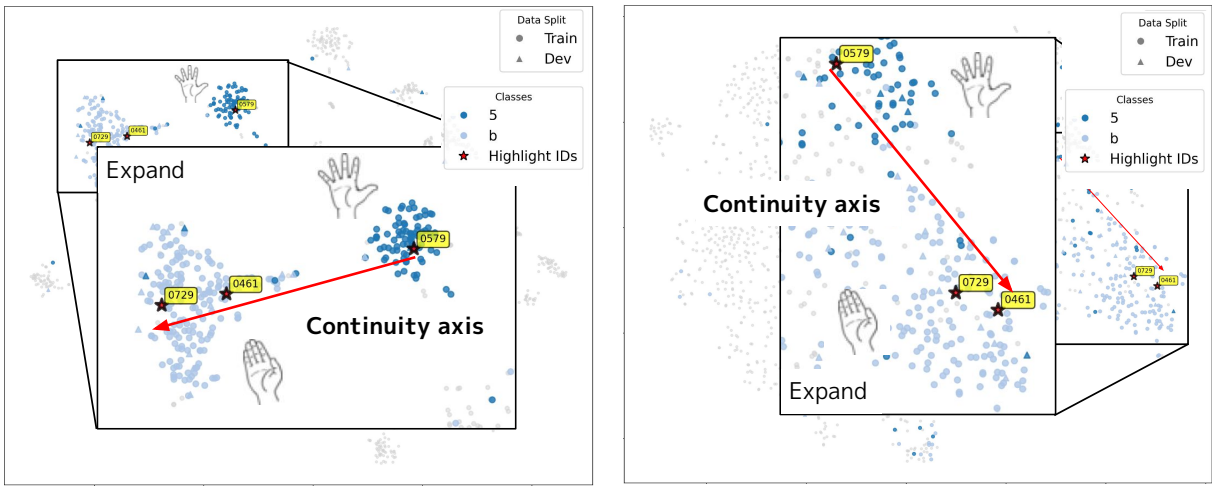
Table 3: Continuity Ranking Accuracy (%) on the JSL Ordered Triplet Dataset. HS: Handshape; Pre/Post: Pre-change/Post-change.

6.2. Phonological Component Classification

Having established that our method successfully learns embeddings that preserve continuous phonological structure, we now evaluate whether these embeddings also maintain discriminative power for the standard classification task. The results on the development set are summarized in Table 4. Notably, the PU-AUC model achieves comparable or better classification accuracy on most components despite being optimized for ranking rather than classification. For instance, PU-AUC shows improvements on dominant hand location (85.09% vs. 78.88%) and non-dominant hand movement (90.68% vs. 83.85%). The performance is mixed on some components, with the CE baseline performing slightly better on dominant hand movement (62.73% vs. 59.01%). These results demonstrate that our approach not only captures continuous phonological structure, but also maintains competitive discriminative ability for categorical prediction.

6.3. Discussion

Qualitative Analysis of Learned Embeddings To qualitatively evaluate how well the learned geometric space reflects phonological similarity, we visualize the embeddings via t-SNE (Maaten and Hinton, 2008) (Figure 5). The visualization reveals that the standard ViT-CE classification baseline does not arrange instances according to their physical progression; while it separates the anchor



(a) ViT-CE (Baseline)

(b) ViT-PU-AUC (Ours)

Figure 5: t-SNE visualization of the learned embedding space for the handshape parameter, showing a continuous transition from one handshape ('0579') through an intermediate form ('0729') to another handshape ('0461'). (a) The classification baseline separates classes but does not preserve the continuous ordering. (b) Our PU-AUC method correctly captures the continuous progression, positioning the intermediate form between the two anchor handshapes in the embedding space.

Component	Method	D	ND
HS (Pre)	CE	46.58	66.46
	PU-AUC	**47.83**	**71.43**
HS (Post)	CE	47.83	65.22
	PU-AUC	**50.93**	**70.81**
Location	CE	78.88	87.58
	PU-AUC	**85.09**	**88.82**
Movement	CE	**62.73**	83.85
	PU-AUC	59.01	**90.68**

Table 4: Phonological Component Classification Accuracy (%) on Dev Set. D: Dominant hand; ND: Non-dominant hand; HS: Handshape.

classes, it does not preserve a direct correspondence between geometric distance and phonological similarity (Figure 5a). In contrast, our ViT-PU-AUC model preserves this relationship, creating a continuous trajectory that mirrors the physical movement of the sign (Figure 5b). This result suggests that the ranking objective of PU-AUC optimization enables the model to learn a smooth representation that captures the continuous nature of articulation from labels.

Performance Analysis Across Components The performance of our method varies across different phonological components, which reflects the differing degrees of continuity inherent in each parameter. In the continuity ranking evaluation (Table 3), Movement, which exhibits clear geometric continuity in physical space, shows the strongest performance (96.00% on the dev set). Handshape continuity, which involves multiple articulatory axes (finger extension, flexion, and spread), shows more moderate improvements, particularly for post-change handshape (97.44% on the dev set). Location, which can involve both spatial position and contact distinctions, shows the most challenging case (81.40% on the dev set), likely because contact introduces a boundary within an otherwise continuous spatial parameter. These patterns suggest that the degree of continuity captured by our model reflects the underlying articulatory properties of each phonological dimension.

7. Conclusion

In this paper, we addressed the standard approach of representing sign language phonology with categories. We proposed learning continuous phonological embeddings by reframing the task as a ranking problem via Positive-Unlabeled (PU) AUC optimization. A key contribution of this work is the creation of a new dataset of phonological progressions, which enables the evaluation of continuous representations. Our intrinsic evaluations on this dataset show that our framework encourages the formation of a continuous, linguistically plausible structure from labels, with our embeddings being more effective at capturing the graded nature of phonological transitions than classification baselines. This is further supported by the analysis on the phonological component classification task, which showed that our method main-

tains competitive discriminative performance. This demonstrates that our method successfully learns continuous phonological structure without sacrificing discriminative performance. By providing a method to learn and evaluate continuous phonological representations, we enable more linguistically-grounded models of signed languages.

Ethics and Limitations Statement

This study, while presenting a new direction, has several limitations. First, the JSL Syllable Database (Yawata et al., 2017) is relatively small (1,078 instances) and features only a single native signer, which poses potential constraints on model generalizability and the reliability of our evaluation. The single-signer limitation means our model learns phonological distinctions specific to one signing style, and generalization to other signers remains to be validated. Larger and more diverse datasets with multiple signers are needed to further validate our findings. Second, the annotations for the triplet dataset were provided by two trained annotators. To further validate the consistency of these perceptual judgments and create more robust benchmarks, future work should involve a larger and more diverse group of annotators. Third, our current work focuses on isolated signs. An important next step is to extend this model to the context of continuous sign language, where phenomena like co-articulation dynamically modulate these continuous representations. Investigating these effects is essential for understanding the dynamics of sign language phonology.

Data Source and Consent The JSL Syllable Database used in this study was constructed in prior research (Yawata et al., 2017). This dataset is available upon request. For this research, we contacted the original authors directly and obtained their permission to use the dataset for our research purposes. The use of the data adheres to the scope of the informed consent originally provided by the signer.

Societal Impact The goal of this research is to develop models that more faithfully capture the linguistic properties of sign language. In the long term, this could contribute to the improvement of sign language recognition and translation technologies, potentially benefiting the Deaf and Hard of Hearing community.

Use of AI Tools This research employed AI-based tools to support code development and to enhance the linguistic quality of the manuscript.

Bibliographical References

Jessa Bekker and Jesse Davis. 2020. Learning from positive and unlabeled data: A survey. *Machine learning*, 109(4):719–760.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703.

Aaron Defazio, Xingyu Alice Yang, Ahmed Khaled, Konstantin Mishchenko, Harsh Mehta, and Ashok Cutkosky. 2024. The road less scheduled. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Susan Goldin-Meadow and Diane Brentari. 2017. Gesture, sign, and language: The coming of age of sign language and gesture studies. *Behavioral and brain sciences*.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. LLMs are Good Sign Language Translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18362–18372.

Thomas Hanke. 2004. HamNoSys – representing sign language data in language resources and language processing contexts. In *Proceedings of the LREC2004 Workshop on the Representation and Processing of Sign Languages: From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication*. European Language Resources Association (ELRA).

Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. 2020. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8129–8138.

Lee Kezar, Jesse Thomason, and Zed Sehyr. 2023. Improving Sign Recognition with Phonology. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2732–2737, Dubrovnik, Croatia. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

D. McKee, R. McKee, S. Pivac Alexander, L. Pivac, and M Vale. 2011. Online dictionary of New Zealand Sign Language. https://www.nzsl.nz/.

Tomoya Sakai, Gang Niu, and Masashi Sugiyama. 2018. Semi-supervised AUC optimization based on positive-unlabeled learning. *Machine Learning*, 107(4):767–794.

Zed Sevcikova Sehyr, Naomi Caselli, Ariel M Cohen-Goldberg, and Karen Emmorey. 2021. The ASL-LEX 2.0 Project: A Database of Lexical and Phonological Properties for 2,723 Signs in American Sign Language. *The Journal of Deaf Studies and Deaf Education*.

William C. Stokoe. 1960. *Sign Language Structure: An Outline of the Visual Communication System of the American Deaf*, volume 8 of *Studies in Linguistics, Occasional Papers*. University of Buffalo, Buffalo, NY.

Valerie Sutton. 1990. Lessons in sign writing. SignWriting.

Federico Tavella, Viktor Schlegel, Marta Romeo, Aphrodite Galata, and Angelo Cangelosi. 2022. WLASL-LEX: a dataset for recognising phonological properties in American Sign Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 453–463, Dublin, Ireland. Association for Computational Linguistics.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. VideoMAE V2: Scaling Video Masked Autoencoders With Dual Masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560.

Satoshi Yawata and Makoto Miwa and Yutaka Sasaki and Daisuke Hara. 2017. *JSL Syllable Database*. Asian Federation of Natural Language Processing.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Ronglai Zuo, Fangyun Wei, and Brian Mak. 2023. Natural Language-Assisted Sign Language Recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14890–14900.

Language Resource References

Jundai Inoue and Daisuke Hara and Makoto Miwa. 2026. *JSL Ordered Triplet Dataset*. To be released upon publication. A dataset of phonological progressions for evaluating continuous representations of Japanese Sign Language.

Satoshi Yawata and Makoto Miwa and Yutaka Sasaki and Daisuke Hara. 2017. *JSL Syllable Database*. Asian Federation of Natural Language Processing.

A. Hyperparameter Settings

Full hyperparameter settings used in all experiments are listed in Table 5. Early stopping is applied based on classification accuracy on the development set.

Table 5: Full hyperparameter settings. SF: Schedule-Free.

Parameter	**Value**
Model Architecture	
Video ViT Backbone	ViT-B/16-224
Embedding dimension	128
Data Augmentation	
Input resolution	224×224
Sampled frames	16
Repeated Aug. (Hoffer et al., 2020)	2
Color Jitter	0.4
RandAugment (Cubuk et al., 2020)	✓
Random Erase (Zhong et al., 2020)	✓
Optimization	
Optimizer	SF-AdamW (Defazio et al., 2024)
Learning rate	1.0×10^{-4}
Weight decay	0.0
Layer decay	1.0
Batch size	2
Gradient accumulation	16 steps
Total epochs	300
DropPath	0.0
Early stopping (patience)	30 epochs
PU-AUC Specific	
Class prior π_c	$1/\|C\|$ (uniform)