Self-Supervised Learning on Lingual Ultrasound Video Encodes Clinically Meaningful Articulatory Structure of Rhotic Production in Residual Speech Sound Disorder

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Speech sound disorder (SSD) affects many school-aged children and can lead to long-term academic, social, and mental health difficulties. Speech-language pathologists use ultrasound, a non-invasive imaging modality, to visualize tongue motion during speech and provide biofeedback treatment. The American English rhotic (r sound) is particularly challenging, as distinct articulatory strategies can produce the same perceived sound. Label-efficient self-supervised learning systems are increasingly being developed to automate scalable and objective SSD assessment. However, existing work has primarily focused on phoneme-level classification and provides limited insight into *within-phoneme* articulatory variations. In this paper, we present the first self-supervised learning framework that focuses on *within-phoneme* articulatory structure of American English rhotic production from lingual ultrasound video of children with residual SSD. We further introduce speaker adversarial regularization to reduce inter-speaker variability while preserving articulatory structure in the learned representations. Our results on the recently curated PERCEPT-US multimodal corpus demonstrate that self-supervised learning without task specific fine-tuning yields embeddings that accurately discriminate between perceptually accurate and inaccurate rhotic productions (mean accuracy: 0.974, mean MCC: 0.941) and between articulatory subtypes of perceptually accurate rhotics (mean accuracy: 0.961, mean MCC: 0.897), namely retroflex and bunched configurations. Code is available at Anonymous Repo.

Keywords: AI assisted speech therapy · self-supervised representation learning · lingual ultrasound video · speech sound disorder

1 Introduction

Speech sound disorders (SSDs) affect up to 10% of preschool and school-age children and account for a substantial proportion of pediatric communication disorders [3, 10]. Residual SSD (RSSD) refers to speech errors that persist beyond the typical developmental window [15], and are associated with long-term academic, social, and mental health consequences [20, 22]. Tongue ultrasound

provides non-invasive visualization of tongue motion via a probe placed beneath the chin, enabling both the clinician and the child to observe the produced tongue shape and offer articulatory cues to support adjustments [6]. Despite its clinical value, ultrasound analysis is labor intensive and requires expert annotation of complex tongue motion, limiting its scalability in routine workflows.

Self-supervised learning (SSL) offers a label-efficient framework that can leverage unlabeled ultrasound video to learn clinically meaningful articulatory features, reducing the reliance on manual annotation [1]. Existing lingual ultrasound SSL methods have demonstrated strength in phoneme-level classification [4, 23, 25], predominantly with reconstruction-based models such as video masked autoencoders (VideoMAE; [17]). However, these current approaches still fail to jointly address the following three challenges in SSL for lingual ultrasound.

First, providing articulatory feedback for the American English *rhotic* is particularly challenging as distinct tongue configurations can produce perceptually equivalent sounds [3, 6]. Most ultrasound corpora lack annotations for within-rhotic subtypes [5], so whether SSL can recover *within-phoneme* structure, specifically *within-rhotic*, is unexplored [1]. Second, inter-speaker variability [8, 13, 24] arises from anatomical differences and speaking style, and is elevated in children still developing speech-motor control. Current SSL pipelines [4, 23, 25] do not control for this variation, obscuring articulatory structure in the learned embeddings and reducing interpretability. Third, existing methods emphasize pixel-level recovery, which is sensitive to imaging noise and captures insufficient high-level articulatory structure [11]. Contrastive learning has seen limited adoption due to the difficulty of defining reliable negatives for highly variable articulatory data [4, 11]. While never applied to tongue ultrasound, non-contrastive joint-embedding frameworks such as BYOL [9] and RESA [21] sidestep both limitations and have shown strong discriminative properties in other domains [16, 19, 21], suggesting a promising direction for articulatory diagnosis.

We address these gaps by introducing a joint-embedding SSL framework to learn clinically relevant articulatory representations from lingual ultrasound video, termed **BYOL** with **V**iViT and **S**peaker-adversarial network (BYOL-VS). The method utilizes the recent PERCEPT-US corpus [5], which is twice larger than the widely used Ultrasuite [7] and uniquely provides expert labeled within-rhotic categories. It further integrates an adversarial network that suppresses speaker-specific variation in the embeddings to better preserve high-level articulatory information. The key contributions of our work are as follows:

1. To the best of our knowledge, we present the first SSL study of within-phoneme articulatory structure in the American English rhotic utilizing the PERCEPT-US corpus, the largest child lingual ultrasound dataset to date.
2. We propose BYOL-VS, an SSL framework that employs spatiotemporal encoding and integrates adversarial training to reduce inter-speaker variability, with minimal computational overhead relative to BYOL.
3. Our method substantially outperforms state-of-the-art baselines on perceptual accuracy and rhotic subtype classification and yields interpretable representations that are aligned to expert-labeled tongue shapes.

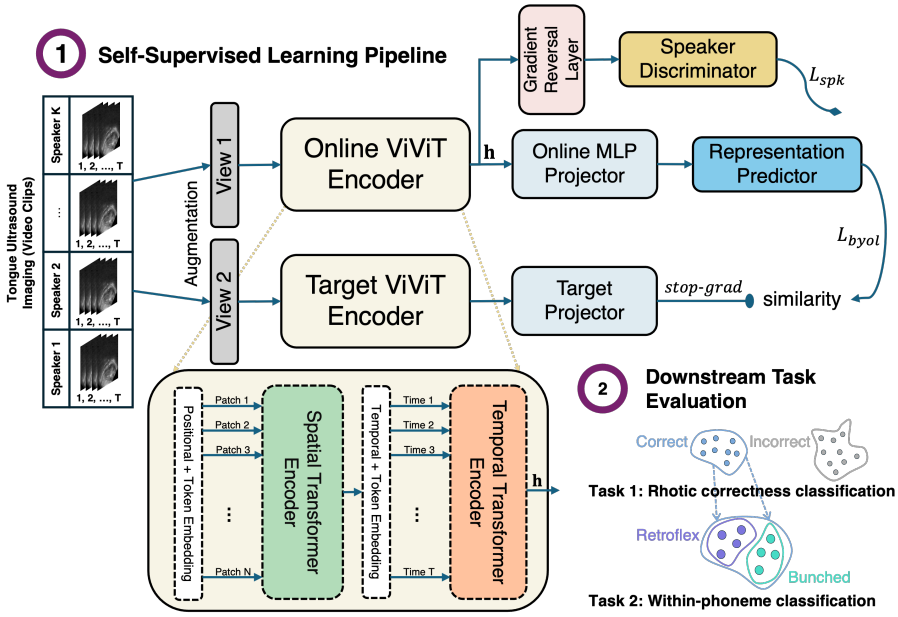

Fig. 1. Overview of BYOL-VS, the proposed joint-embedding SSL pipeline for lingual ultrasound video. An adversarial speaker classifier is attached to the online encoder to suppress speaker-specific information in the learned representations, $\mathbf{h}$. After training, all components except the online ViViT encoder are discarded, and $\mathbf{h}$ is used for two downstream tasks: perceptual accuracy classification and rhotic subtype classification.

2 Method

Objective. We aim to learn representations from lingual ultrasound video that capture clinically relevant *within-phoneme* articulatory structure. Prior work has primarily used supervised learning [14] or self-supervised pretraining followed by supervised fine-tuning [4, 23, 25] to predict phoneme-level labels. In contrast, we demonstrate that SSL *without* task-specific fine-tuning yields embeddings that reliably distinguish both perceptual accuracy and within-phoneme variation, namely bunched and retroflex configurations. Formally, we consider lingual ultrasound clips from K speakers (see Section 3) with each clip represented as $\mathbf{x} \in \mathbb{R}^{T \times H \times W}$, where T is the number of frames and $H \times W$ the spatial resolution.

Self-Supervised Learning of Lingual Ultrasound Video. The proposed SSL pipeline is illustrated in Figure 1. Our framework follows the Bootstrap Your Own Latent (BYOL; [9]) paradigm, which learns view-invariant representations by aligning embeddings of two augmented views of the same input in a joint-embedding space. Unlike reconstruction-based masked modeling or contrastive approaches, it learns predictive features without pixel-level reconstruction or negative samples. It comprises an online network with encoder f_θ, projection

head g_θ, and predictor q_θ, together with a target network with encoder f_ξ and projection head g_ξ. The parameter sets of the online and target networks are denoted by θ and ξ respectively. The ultrasound video input $\mathbf{x}$ undergoes two data augmentations $t_1, t_2 \sim \mathcal{T}$ that generate paired views $\mathbf{v}_1 = t_1(\mathbf{x})$ and $\mathbf{v}_2 = t_2(\mathbf{x})$. The lingual ultrasound appropriate augmentations are described at the end of Section 2. Each view is processed by a Video Vision Transformer (ViViT; [2]) encoder, producing latent representations $\mathbf{h}_i = f_\theta(\mathbf{v}_i)$ and $\mathbf{h}'_j = f_\xi(\mathbf{v}_j)$ for the online and target branches respectively, where $(i, j) \in \{(1, 2), (2, 1)\}$ denotes the two augmented views. These representations are mapped to a projection space $\mathbf{z}_i = g_\theta(\mathbf{h}_i)$ and $\mathbf{z}'_j = g_\xi(\mathbf{h}'_j)$, with gradients stopped for the target branch. The online predictor q_θ is trained to predict the target network's latent representation by minimizing their mean squared error given by

$$\mathcal{L}_{\text{BYOL}} = \frac{1}{B} \sum_{b=1}^{B} \left\| \bar{q}_\theta(\mathbf{z}_i^{(b)}) - \text{sg}(\bar{\mathbf{z}}_j'^{(b)}) \right\|_2^2, \qquad (i, j) \in \{(1, 2), (2, 1)\}, \tag{1}$$

where $b \in \{1, \ldots, B\}$ indexes clips in the batch, $(i, j) \in \{(1, 2), (2, 1)\}$ denote the augmented views, and $\bar{\mathbf{u}} = \mathbf{u}/\|\mathbf{u}\|_2$ denotes ℓ_2 normalization. Due to the stop gradient (sg) placed in the target network, only the online branch that is asymmetric from the target branch is directly optimized via backpropagation. The target network parameters are updated via an exponential moving average $\xi \leftarrow \tau\xi + (1 - \tau)\theta$, with momentum coefficient $\tau = 0.99$.

Video Vision Transformer for Spatiotemporal Encoding. We employ a ViViT [2] as the spatiotemporal encoder. Each frame is partitioned into non-overlapping $P \times P$ patches, yielding $N = (H/P)(W/P)$ patches per frame, which are linearly projected into D-dimensional token embeddings. For each frame at timepoint t, a learnable spatial class token $\mathbf{c}_s \in \mathbb{R}^D$ is prepended to the patch embeddings $\mathbf{E}^{(t)} \in \mathbb{R}^{N \times D}$, and a spatial Transformer encoder is applied across the resulting $(N + 1)$ tokens to produce

$$\mathbf{S}^{(t)} = \text{Transformer}_{\text{space}}\big([\mathbf{c}_s; \mathbf{E}^{(t)}]\big) \in \mathbb{R}^{(N+1) \times D}. \tag{2}$$

The spatial class token representations $\mathbf{S}_0^{(t)} \in \mathbb{R}^D$ are stacked across timepoints $t = 1, 2, \ldots, T$ and prepended with a temporal class token $\mathbf{c}_t \in \mathbb{R}^D$. This sequence is processed by a second Transformer encoder to model temporal dynamics,

$$\mathbf{H} = \text{Transformer}_{\text{time}}\big([\mathbf{c}_t; \mathbf{S}_0^{(1)}, \ldots, \mathbf{S}_0^{(T)}]\big) \in \mathbb{R}^{(T+1) \times D}. \tag{3}$$

The final video representation is the temporal class token representation $\mathbf{h} = \mathbf{H}_0 \in \mathbb{R}^D$, which provides a spatiotemporal summary of all T frames.

Adversarial Suppression of Speaker Information. To suppress speaker-specific information, we attach an adversarial speaker classifier to the online encoder, which predicts the speaker identity $s_b \in \{1, \ldots, K\}$ from the representation $\mathbf{h}$. A gradient reversal layer is placed between $\mathbf{h}$ and the classifier so

that during backpropagation, the encoder receives the *negated gradient* scaled by λ_{grl}, a hyperparameter [8, 13, 24]. The overall objective becomes

$$\mathcal{L} = \mathcal{L}_{\text{BYOL}} + \lambda_{\text{grl}}\,\mathcal{L}_{\text{spk}}, \quad \mathcal{L}_{\text{spk}} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{K} y_k^{(b)} \log \hat{y}_k^{(b)}, \tag{4}$$

where $y_k^{(b)} = \mathbf{1}\{s_b = k\}$ and $\hat{y}_k^{(b)}$ is the predicted softmax probability. Evidently, the classifier is optimized to minimize $\mathcal{L}_{\text{spk}}$, while the encoder receives the *reversed gradient*. Thus, the online encoder update can be described by

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}_{\text{BYOL}}}{\partial \theta} - \lambda_{\text{grl}} \frac{\partial \mathcal{L}_{\text{spk}}}{\partial \theta}, \tag{5}$$

which minimizes the MSE loss (Eq. 1) while discouraging speaker-discriminative information in $\mathbf{h}$. After training, all components except the online ViViT encoder are discarded, and its output $\mathbf{h}$ is used for downstream classification tasks.

Lingual Ultrasound Appropriate Data Augmentations. To model variability from tongue motion, probe orientation, and imaging quality [18], we apply anatomically constrained perturbations consisting of rotations $\theta \in [-10°, 10°]$, translations up to 10% of the image size, and isotropic scaling $s \in [0.9, 1.1]$. Small perspective transformation with distortion scale 0.12 is applied with 0.25 probability to simulate variations in probe angle. Non-rigid tongue motion and compression are modeled using elastic deformations with strength $\alpha \in \{2.0, 5.0\}$ and smoothing $\sigma \in \{0.5, 0.7\}$. Variability across ultrasound machines and acquisition settings is simulated through speckle noise with $\sigma \in [0, 0.08]$, brightness and contrast jitter with factors $\gamma_b, \gamma_c \in [0.8, 1.2]$, and Gaussian blur with $\sigma \in [0.5, 1.2]$. Flips are excluded as all recordings share a consistent probe orientation; flips create anatomically implausible tongue configurations and have been shown to degrade performance in ultrasound analysis [18].

3 Experiments and Results

Dataset. Experiments are conducted using PERCEPT-US [5], the largest American English child speech corpus with lingual ultrasound video data. Images were obtained during ultrasound biofeedback clinical trials focused on children with RSSD. The data comprise recordings from 70 child speakers (ages 8–17) totaling 4,370 utterances. Each clip is annotated by experts with tongue shape categories used in clinical practice. Of these, 3,054 clips are labeled as perceptually accurate and 1,316 as perceptually inaccurate. Among the perceptually accurate rhotic productions, 1,722 are labeled as *bunched*, 667 as *retroflex*, and 665 as ambiguous due to articulatory uncertainty or annotator disagreement. Figure 2 shows two ultrasound frames from each category. Our work is the first SSL study to use PERCEPT-US for articulatory representation learning.

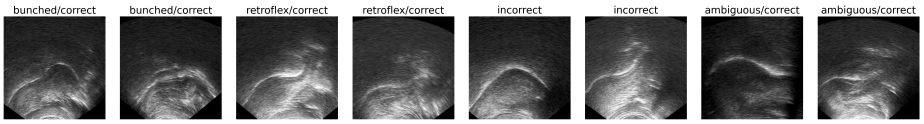

Fig. 2. Expert-labeled ultrasound frames showing perceptually accurate rhotics with bunched and retroflex tongue shapes, perceptually inaccurate productions, and perceptually accurate productions with ambiguous tongue shape.

Evaluation Metrics. Five-fold speaker-disjoint cross-validation (CV) is utilized to assess generalization to unseen speakers. Classification performance is evaluated using a k-nearest neighbor (kNN) probe with $k = 5$, reported as accuracy and Matthews correlation coefficient (MCC), with the mean and standard deviation (SD) calculated over the CV folds. MCC accounts for class imbalance. Uniform Manifold Approximation and Projection (UMAP; [12]) is used for qualitative analysis of clustered embedding structure. We benchmark three state-of-the-art SSL baselines: BYOL (2020; [9]), VideoMAE (2022; [17]), RESA (2025; [21]) which encourages clustering in the encodings, and the proposed BYOL-VS.

Training Details. All models are trained on a single NVIDIA L40S GPU (48 GB VRAM) with 20 Intel Xeon Platinum 8592+ CPU cores. Each clip contains 12 frames resized to 224×224, utilizes 16×16 spatial patch size, and trained for 200 epochs with cosine learning rate decay to 10^{-6} and 20-epoch linear warmup. The AdamW optimizer is used with a weight decay of 0.0001. BYOL uses a batch size of 16 and learning rate 2×10^{-4}. BYOL-VS adopts the same settings with a speaker-adversarial module ($\lambda_{\text{grl}} = 0.1$) and a two layer MLP speaker classifier of hidden size 256. RESA uses a batch size of 16, learning rate 1.5×10^{-4}, and weight decay 10^{-4}. VideoMAE uses a tubelet size of 2, batch size 32, mask ratio 0.9, and learning rate 1×10^{-4}. For fair comparison, BYOL, RESA, and BYOL-VS share the same ViViT backbone with 4 spatial and 4 temporal transformer layers, while VideoMAE uses a ViT encoder following previous works [4, 23, 25].

Quantitative and Qualitative Results. Table 1 summarizes performance on the two downstream classification tasks under speaker-disjoint cross validation. The first task distinguishes perceptually accurate from inaccurate productions, while the second discriminates within-rhotic tongue shapes (retroflex vs. bunched). Compared to VideoMAE, the joint-embedding methods (RESA, BYOL, BYOL-VS) yield consistent improvements of at least 0.096 in mean accuracy and 0.194 in mean MCC for classifying perceptual accuracy, and 0.081 in mean accuracy and 0.189 in mean MCC for classifying within-rhotic tongue shapes. BYOL-VS outperforms RESA and BYOL on both tasks. Table 2 examines the impact of varying λ_{grl}. Increasing the speaker-adversarial term up to $\lambda_{\text{grl}} = 1$ yields higher mean accuracy and MCC on both tasks compared to that of $\lambda_{\text{grl}} = 0$, with best results at $\lambda_{\text{grl}} = 0.1$, declining in performance for subtype discrimination when $\lambda_{\text{grl}} = 5$. This improvement suggests that moderate suppression of speaker-related noise substantially benefits downstream performance.

Table 1. Speaker-disjoint 5-fold CV performance (mean ± SD) evaluated using a k-nearest neighbor (kNN) probe ($k = 5$), without fine-tuning the encoder.

Method	Correct vs. Incorrect		Retroflex vs. Bunched	
	Accuracy	MCC	Accuracy	MCC
VideoMAE [17]	0.808 ± 0.022	0.571 ± 0.068	0.837 ± 0.019	0.580 ± 0.040
RESA [21]	0.916 ± 0.023	0.765 ± 0.061	0.928 ± 0.071	0.769 ± 0.049
BYOL [9]	0.904 ± 0.033	0.799 ± 0.087	0.918 ± 0.031	0.802 ± 0.065
BYOL-VS (ours)	$\mathbf{0.974} \pm \mathbf{0.019}$	$\mathbf{0.941} \pm \mathbf{0.044}$	$\mathbf{0.961} \pm \mathbf{0.013}$	$\mathbf{0.897} \pm \mathbf{0.031}$

Table 2. Ablation study for BYOL-VS showing the effect of the adversarial weight λ_{grl} on the two downstream tasks. Other than λ_{grl}, the hyperparameters are kept identical.

Setting	Correct vs. Incorrect		Retroflex vs. Bunched	
	Accuracy	MCC	Accuracy	MCC
$\lambda_{\text{grl}} = 0$	0.904 ± 0.033	0.799 ± 0.087	0.918 ± 0.031	0.802 ± 0.065
$\lambda_{\text{grl}} = 0.1$	$\mathbf{0.974} \pm \mathbf{0.019}$	$\mathbf{0.941} \pm \mathbf{0.044}$	$\mathbf{0.961} \pm \mathbf{0.013}$	$\mathbf{0.897} \pm \mathbf{0.031}$
$\lambda_{\text{grl}} = 0.2$	0.968 ± 0.014	0.929 ± 0.031	0.956 ± 0.019	0.891 ± 0.043
$\lambda_{\text{grl}} = 1.0$	0.952 ± 0.035	0.896 ± 0.070	0.943 ± 0.029	0.852 ± 0.082
$\lambda_{\text{grl}} = 5.0$	0.924 ± 0.050	0.825 ± 0.125	0.905 ± 0.076	0.761 ± 0.176

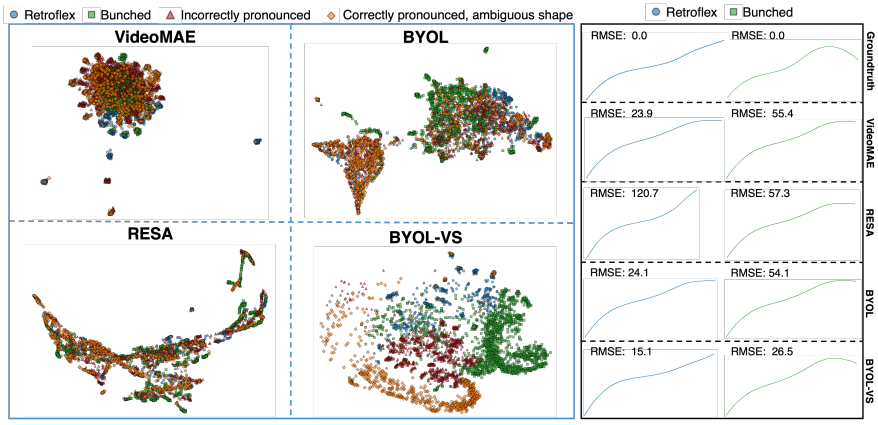

Fig. 3. Qualitative comparison of methods. Left: UMAP visualization of SSL video representations (with 30 UMAP neighbors). Right: Mean tongue contours of the central frame obtained by spectral clustering the embeddings of perceptually accurate clips. Clusters derived from BYOL-VS best recover the correct tongue shape geometry.

Figure 3 provides complementary qualitative evidence. For the qualitative analyses, all models were refit on the full dataset using the best set of hyperparameters selected from the 5-fold CV. On the left, UMAP projections of final video-level encoder representations labeled with ground truth categories show that BYOL-VS exhibits more visually separable clusters than competing methods. On the right, we evaluate whether this structure reflects articulatory geometry. Instead of labeling the embeddings with true articulatory classes (retroflex,

Table 3. Mean ± SD of training time and peak VRAM usage per CV fold.

Method	Runtime (hours)	VRAM (GB)
VideoMAE [17]	3.22 ± 0.22	10.51 ± 0.59
RESA [21]	11.56 ± 0.15	9.75 ± 0.07
BYOL [9]	2.56 ± 0.11	5.75 ± 0.15
BYOL-VS (ours)	3.09 ± 0.15	7.16 ± 0.15

bunched), unsupervised spectral clustering with two clusters is first applied to the embeddings of clips predicted as perceptually accurate. For each cluster, we compute the mean tongue contour from the central frame using 100 expert-traced coordinates and root mean squared error (RMSE) as the pointwise deviation from the ground truth contours. BYOL-VS achieves the lowest RMSE and yields contours that most closely match the ground truth tongue shapes, suggesting its representations better preserve articulatory geometry. RESA produces interpretable contours but with substantially higher RMSE, consistent with less accurate separation in the embeddings. BYOL and VideoMAE show limited visual distinction between the two configurations, with higher RMSE than that of BYOL-VS for both shapes. While reconstruction-based objectives emphasize low-level detail and can limit higher-level abstraction needed for classification [11], our results suggest that joint-embedding methods are similarly affected when domain-specific variability, such as inter-speaker differences, is not explicitly controlled. With adversarial training, BYOL-VS recovers known structures in perceptually accurate productions, indicating the suppression of noise improved the learning of articulatory information.

Computational Efficiency. The computational cost is summarized in Table 3. BYOL is the most efficient in both runtime and memory, VideoMAE requires higher VRAM, and RESA is the slowest. BYOL-VS introduces an overhead relative to BYOL but remains more efficient than VideoMAE and RESA. Given the substantial improvement in performance, this modest overhead is worthwhile.

4 Conclusion

We introduce BYOL-VS, a speaker-adversarial SSL framework for learning clinically relevant within-rhotic articulatory structure from lingual ultrasound without task-specific fine-tuning. We demonstrate that joint-embedding SSL method can well encode both perceptual accuracy and rhotic subtype structure when speaker-specific variation is controlled. Future work will extend the analysis to perceptually inaccurate and ambiguous productions to derive standardized articulatory cues that can improve consistency and effectiveness in clinical intervention across phoneme classes. In addition, the framework will be expanded to include a broader set of phoneme-level articulatory categories and to incorporate synchronized acoustic signals for multimodal representation learning.

Bibliography

[1] Al Ani, S., Cleland, J., Zoha, A.: Deep learning in ultrasound tongue imaging: a systematic review toward automated detection of speech sound disorders. Frontiers in artificial intelligence **8**, 1631134 (2025)

[2] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021)

[3] Byun, T.M., Hitchcock, E.R., Swartz, M.T.: Retroflex versus bunched in treatment for rhotic misarticulation: Evidence from ultrasound biofeedback intervention. Journal of Speech, Language, and Hearing Research **57**(6), 2116–2130 (2014)

[4] Dan, X., Xu, K., Zhou, Y., Yang, C., Chen, Y., Dou, Y., Yang, C.: Spatio-temporal masked autoencoder-based phonetic segments classification from ultrasound. Speech Communication **169**, 103186 (2025)

[5] Eads, A., Kabakoff, H., Benway, N., Hitchcock, E., Preston, J., McAllister, T.: PERCEPT-US: A multimodal american english child speech corpus specialized for articulatory feedback. In: Proc. Interspeech 2025. pp. 2805–2809 (2025)

[6] Eads, A., Kabakoff, H., King, H., Preston, J.L., McAllister, T.: An articulatory analysis of american english rhotics in children with and without a history of residual speech sound disorder. Journal of Speech, Language, and Hearing Research **67**(11), 4246–4263 (2024)

[7] Eshky, A., Ribeiro, M.S., Cleland, J., Richmond, K., Roxburgh, Z., Scobbie, J.M., Wrench, A.: UltraSuite: A Repository of Ultrasound and Acoustic Data from Child Speech Therapy Sessions. In: Interspeech 2018. pp. 1888–1892 (2018)

[8] Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)

[9] Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)

[10] Hitchcock, E.R., Harel, D., Byun, T.M.: Social, emotional, and academic impact of residual speech errors in school-aged children: A survey study. In: Seminars in speech and language. vol. 36, pp. 283–294. Thieme Medical Publishers (2015)

[11] Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J.: Self-supervised learning: Generative or contrastive. IEEE Transactions on Knowledge and Data Engineering **35**(1), 857–876 (2023)

[12] McInnes, L., Healy, J., Saul, N., Grossberger, L.: Umap: Uniform manifold approximation and projection. The Journal of Open Source Software **3**(29), 861 (2018)

[13] Meng, Z., Li, J., Chen, Z., Zhao, Y., Mazalov, V., Gong, Y., Juang, B.H.: Speaker-invariant training via adversarial learning. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5969–5973. IEEE (2018)

[14] Ribeiro, M.S., Cleland, J., Eshky, A., Richmond, K., Renals, S.: Exploiting ultrasound tongue imaging for the automatic detection of speech articulation errors. Speech Communication **128**, 24–34 (2021)

[15] Spencer, C., Vannest, J., Maas, E., Preston, J.L., Redle, E., Maloney, T., Boyce, S.: Neuroimaging of the syllable repetition task in children with residual speech sound disorder. Journal of Speech, Language, and Hearing Research **64**(6S), 2223–2233 (2021)

[16] Tian, Y., Chen, X., Ganguli, S.: Understanding self-supervised learning dynamics without contrastive pairs. In: International Conference on Machine Learning. pp. 10268–10278. PMLR (2021)

[17] Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35**, 10078–10093 (2022)

[18] Tupper, A., Gagné, C.: Revisiting data augmentation for ultrasound images. Transactions on Machine Learning Research (2025)

[19] Van Assel, H., Ibrahim, M., Biancalani, T., Regev, A., Balestriero, R.: Joint embedding vs reconstruction: Provable benefits of latent space prediction for self supervised learning. arXiv preprint arXiv:2505.12477 (2025)

[20] Vick, J.C., Campbell, T.F., Shriberg, L.D., Green, J.R., Truemper, K., Rusiewicz, H.L., Moore, C.A.: Data-driven subclassification of speech sound disorders in preschool children. Journal of Speech, Language, and Hearing Research **57**(6), 2033–2050 (2014)

[21] Weng, X., An, J., Ma, X., Qi, B., Luo, J., Yang, X., Dong, J.S., Huang, L.: Clustering properties of self-supervised learning. Forty-second International Conference on Machine Learning (2025)

[22] Wren, Y., Pagnamenta, E., Orchard, F., Peters, T.J., Emond, A., Northstone, K., Miller, L.L., Roulstone, S.: Social, emotional and behavioural difficulties associated with persistent speech disorder in children: A prospective population study. JCPP advances **3**(1), e12126 (2023)

[23] Xu, K., You, K., Zhu, B., Feng, M., Feng, D., Yang, C.: Masked modeling-based ultrasound image classification via self-supervised learning. IEEE Open Journal of Engineering in Medicine and Biology **5**, 226–237 (2024)

[24] Yin, Y., Huang, B., Wu, Y., Soleymani, M.: Speaker-invariant adversarial domain adaptation for emotion recognition. In: Proceedings of the 2020 International Conference on Multimodal Interaction. pp. 481–490 (2020)

[25] You, K., Liu, B., Xu, K., Xiong, Y., Xu, Q., Feng, M., Csapó, T.G., Zhu, B.: Raw ultrasound-based phonetic segments classification via mask modeling. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)