PULSE: A Principled Framework for Multimodal Speech Sound Disorder Assessment Using Ultrasound Tongue Imaging

Varun Shijo, Anarghya Das, Wei Bo, Shuwei Hou, Ling-Yu Guo, Jun Xia, Wenyao Xu
University at Buffalo, Buffalo, NY, USA
{varunshi, anarghya, weibo, shuweiho, lingyugu, junxia, wenyaoxu}@buffalo.edu

***Abstract*—Ultrasound tongue imaging (UTI) has demonstrated clinical effectiveness for speech sound disorder (SSD) assessment and biofeedback therapy, yet its adoption in routine practice remains limited. While recent deep learning approaches operate directly on ultrasound frames, a systematic review identifies interpretability, speaker generalization, and clinical translation as critical unmet needs, leaving interpretable analysis as a key barrier to adoption. To address these gaps, we propose PULSE, a principled framework for feature discovery and selection in multimodal SSD assessment. PULSE extracts interpretable articulatory features from keypoint trajectories over tongue pose sequences, organizes them into semantically meaningful groups (spatial, shape, kinematic), and evaluates each group's contribution through systematic ablation under speaker-independent protocols. This methodology reveals that static features (positional and shape) encode speaker identity rather than production quality, thereby degrading generalization when naively fused with audio. Domain-knowledge-driven group selection resolves this degradation by discarding subject-variant features and retaining only phoneme-variant kinematic features that capture movement dynamics. Validated on the UltraSuite dataset's UXSSD corpus under leave-one-speaker-out evaluation on 9 children with SSD, selective fusion of kinematic ultrasound features with audio achieves AUC 0.725 versus 0.687 for audio alone, correctly reclassifying 68 instances that audio misses with zero losses. These results show that principled, interpretable feature selection yields more speaker-generalizable multimodal SSD assessment than naive fusion.**

***Index Terms*—speech sound disorders, ultrasound tongue imaging, multimodal fusion, interpretable framework**

I. Introduction

Speech sound disorders (SSD) affect 5–8% of children [1], and clinical assessment relies on perceptual judgment by speech-language pathologists (SLPs), which is subjective and cannot capture articulatory detail beyond what can be heard. Ultrasound tongue imaging (UTI) addresses this gap by visualizing tongue position and movement in real time, and ultrasound visual biofeedback has demonstrated effectiveness for persistent SSD [2]. However, interpreting UTI requires manual expert analysis that is time-consuming and impractical in routine clinical settings [3].

Prior work has produced publicly available SSD data with expert pronunciation scores [3], sub-millimeter tongue pose estimation [4], and tongue shape complexity metrics validated on clinical data [5]. However, each was evaluated in isolation and no prior work has leveraged these components for principled, speaker-independent clinical assessment. Deep learning-based approaches [6], [7] operate on raw frames without speaker-independent evaluation or interpretable outputs, and a recent systematic review [8] identifies interpretability, speaker generalization, and clinical translation as the critical unmet needs.

We propose PULSE (Pose-based ULtrasound Speech Evaluation), a framework for translating UTI into automated, interpretable SSD assessment through systematic feature discovery and selection. PULSE extracts interpretable articulatory features from tongue keypoint trajectories on the UltraSuite SSD corpus, organizes them into semantically meaningful groups (spatial, shape, kinematic), and applies systematic ablation to identify which generalize across speakers under leave-one-speaker-out (LOSO) evaluation. Specifically, our contributions are: (1) interpretable articulatory features adapted from movement science and speech morphometrics (temporal binning, kinematic descriptors, DFT contour shape) and visualization tools (*lingual differentiation index*, LDI; *lingual workspace maps*, LWMs) that characterize the motor basis of SSD. (2) A systematic ablation methodology revealing that static ultrasound features encode speaker identity, and domain-knowledge-driven group selection that retains only kinematic features, achieving the best speaker-independent performance. (3) Speaker-independent evaluation showing selective fusion improves over audio alone (AUC 0.725 vs 0.687) and recovers 68 instances that audio misclassifies with zero losses.

II. Related Work

Multimodal UTI Dataset. The UltraSuite repository [3] provides synchronized ultrasound and audio from 58 TD children (UXTD), 8 with SSD (UXSSD), and 20 therapy participants (UPX). To our knowledge, the SLP-annotated pronunciation scores for the SSD subsets (UXSSD, UPX) have not been used for computational assessment, presenting an opportunity for automated, speaker-independent analysis.

Tongue pose estimation and feature validation. Wrench and Balch-Tomes [4] demonstrated that DeepLabCut achieves 0.93 mm MSD on 16 tongue keypoints, matching human expert annotators. Downstream feature extraction and clinical application, however, remain largely unexplored. Kabakoff et al. [5] validated tongue shape metrics (MCI, NINFL), show-

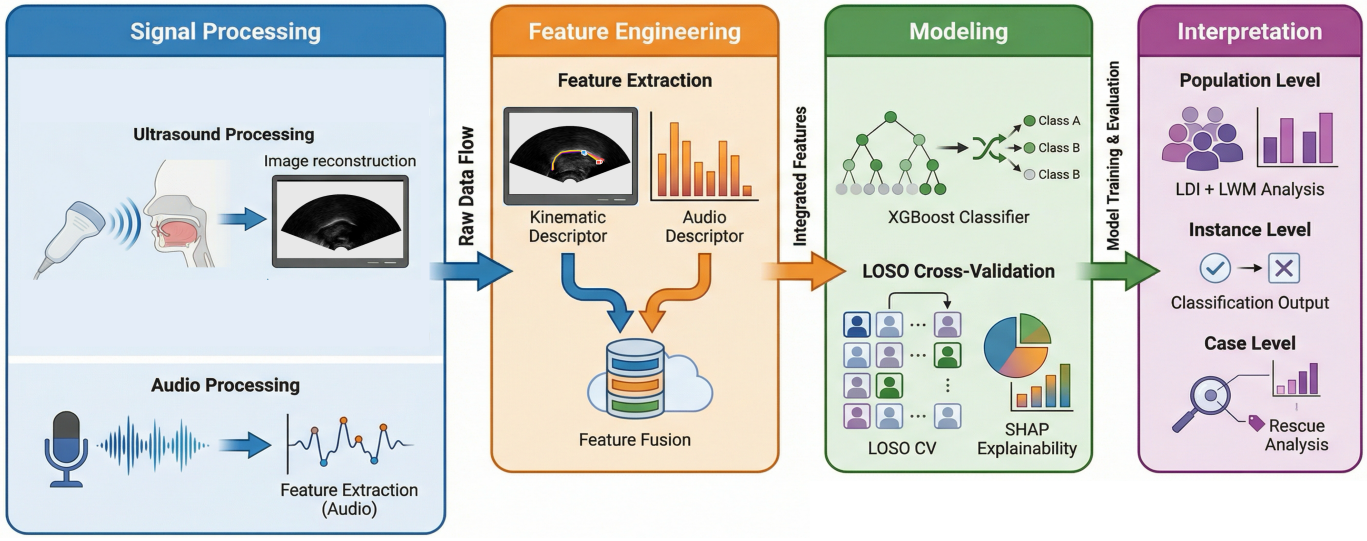


Fig. 1. PULSE framework overview. Signal processing converts raw ultrasound into calibrated tongue poses. Feature engineering extracts kinematic and audio descriptors, with selective fusion excluding static anatomical features that encode speaker identity (Section IV-B). Speaker-independent modeling under LOSO evaluation produces population-level articulatory characterization, instance-level scoring, and case-level evidence for clinical review.

ing NINFL discriminates correct from incorrect /r/ ($p = 0.002$) in 24 children via statistical tests.

Automated approaches. Ribeiro et al. [6] trained a CNN on raw frames fused with MFCCs, achieving 86.6% velar fronting detection on UXSSD with per-speaker Cohen's κ ranging from near-perfect to slight agreement. Such approaches, while effective, use learnt features that are opaque. Furthermore, training on TD speech doesn't account for intra-phoneme production quality variance. Dan et al. [7] applied masked autoencoders to UXTD for 4-way phoneme classification (85.3% accuracy), however, this being a phoneme identification task leaves the disorder severity assessment gap open. Aytutuldu et al. [9] learned UTI-audio embeddings for audio-only deployment, the opposite of our biofeedback setting where the probe is present. A systematic review [8] identifies speaker generalization, interpretability, and clinical translation as the critical open gaps.

Clinical motivation. Cleland et al. [10] showed that children with persistent velar fronting exhibit articulatory collapse detectable only via UTI. This covert contrast scenario motivates UTI biofeedback [2] and interpretable articulatory features that complement audio.

III. Methods

A. Dataset

We use the UltraSuite corpus [3], combining UXSSD (8 children with SSD, ages 6–14) and UPX (speakers with confirmed articulation disorder) for a total of 9 SSD speakers and 1,151 phoneme instances scored by experienced SLPs on a 1–5 scale (/r/, /k/, /g/). Within UPX, scored data come from a single child (04M) recorded across four longitudinal therapy sessions (baseline, mid, post, maintenance), contributing 288 of the 1,151 phoneme instances. A lack of phoneme-level annotations in UXTD precludes a more straightforward formulation of UXTD v/s UXSSD for granular features suitable for realtime interventional biofeedback use. Scores are unevenly distributed across levels (1: 454, 2: 119, 3: 102, 4: 227, 5: 249). Several levels have fewer than 12 instances per speaker, making fine-grained classification infeasible under LOSO with 9 speakers. Working within these constraints we binarize to correct (score = 5) versus impaired (1–4). UXTD (58 TD children, ∼2,603 phonemes) serves as the typically developing reference population for articulatory characterization (Section IV-A) and is not used for classification.

TABLE I
Feature taxonomy. Binned features: 5 temporal windows × {mean, std} per base feature. Aggregates: one value per phoneme.

Group	Base	Total	Type
Ultrasound (from tongue pose keypoints)			
Anatomical†	4	40	0th order (static)
Positional†	4	40	0th order (static)
Shape†	11	110	0th order (static)
Kinematic (binned)	4	40	1st–3rd order
Kinematic (agg.)	8	8	cumulative
US subtotal		*238*	
Audio (from waveform)			
Spectral + formant	45	450	—
Total		**688**	

†Speaker-dependent; excluded in selective fusion.

B. Processing Pipeline

Our pipeline (Fig. 1) processes raw `.ult` files (63 scanlines × 412 pixels in fan-beam polar geometry) through three stages: (1) polar-to-Cartesian reconstruction via bilinear interpolation, producing 883×487 Cartesian frames; (2) DeepLabCut [11] pose estimation using the pretrained US_ResNet50

model [4], producing 16 keypoints per frame (11 tongue surface points from vallecula to tip, plus 5 anatomical landmarks) with exponential confidence smoothing ($\alpha = 0.95$); and (3) rectification from pixel coordinates to mm-space via the known probe geometry (angle $\in [-70°, +70°]$, depth $\in [5, 50]$ mm), yielding Cartesian (x, y) positions in millimeters allowing features to generalize across probe geometries.

C. Feature Extraction

Each phoneme is divided into 5 temporal bins (0–20%, …, 80–100%), following percentage-based temporal normalization common in biomechanical movement analysis [12], with per-bin features computed as the mean and standard deviation. Prior work uses only segment-level or frame-level features, without capturing within-phoneme dynamics. Table I summarizes the full feature set.

Anatomical features (tongue length, height, spread, tip-to-dorsum distance) and positional features (tip and dorsum x, y coordinates) capture static tongue geometry. Shape features include MCI and NINFL [5], spline-fit curvature, and DFT coefficients 1–8 decomposing the tongue contour into frequency components following Dawson et al. [13]. Kinematic features capture movement, including tip and dorsum speed (Savitzky–Golay filter, window = 5, order = 2) and jerk, plus whole-segment path length, displacement, excursion, and normalized jerk—a dimensionless smoothness metric from motor rehabilitation [14]. Audio features (16 kHz, 512-sample FFT, 160-sample hop) include 13 MFCCs with Δ and $\Delta\Delta$, spectral centroid, bandwidth, rolloff, flatness, RMS energy, zero-crossing rate, and formants F1–F3 via LPC.

Multimodal fusion. *Naive* late fusion concatenates all 688 features without curation, while *selective* fusion applies domain-knowledge-driven feature group selection, retaining only audio and the 48 kinematic ultrasound features (1st–3rd order derivatives and cumulative descriptors) and excluding the 190 static features (anatomical, positional, shape) marked † in Table I. This exclusion is motivated by the hypothesis that static tongue geometry encodes speaker identity rather than production quality, as validated by the ablation study in Section IV-B.

D. Classification and Evaluation

The binary task has 249 correct (22%) and 902 impaired (78%) instances. We train XGBoost [15] classifiers (100 trees, max depth 6, learning rate 0.1) with default hyperparameters. For each configuration, a single decision threshold is chosen by Youden's J statistic on the pooled leave-one-speaker-out predictions ($J^* = 0.096$ for selective fusion). AUC and PR-AUC are reported as threshold-independent metrics. We report **GroupStratifiedKFold** (GKF, 4 folds, $H \approx 2$ held-out speakers per fold) and **Leave-One-Speaker-Out** (LOSO, train on 8, test on 1). We use SHAP TreeExplainer [16] for feature analysis.

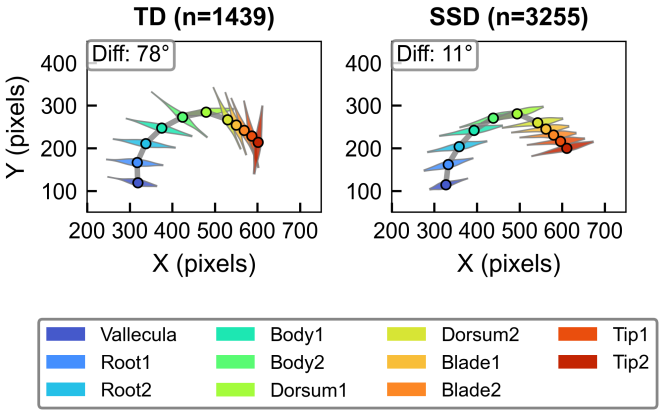


Fig. 2. Spatial variability per tongue keypoint (TD vs SSD). Ellipse orientation shows principal variance direction. TD: LDI = 78°; SSD: LDI = 11°, indicating rigid-body movement.

IV. Results

A. Articulatory Characterization of TD vs SSD

We introduce two visualization methods to characterize articulatory differences from the pose data.

Lingual Differentiation Index (LDI). For each tongue surface keypoint $k \in \{0, \ldots, 10\}$, we compute the first principal component of all mm-space positions, yielding an angle $\theta_k = \arctan(|y_{\text{PC1}}|/|x_{\text{PC1}}|) \in [0°, 90°]$ capturing the principal direction of variance. The *lingual differentiation index* (LDI) is the angular range $\text{LDI} = \max_k \theta_k - \min_k \theta_k$, where high LDI indicates that distinct tongue regions move independently (enabling complex shapes), whereas low LDI indicates the tongue moves as a single rigid body.

TD children exhibit LDI = 78°, reflecting independent control of distinct tongue regions (Fig. 2), whereas SSD children show 11°, indicating rigid-body movement. This metric grades monotonically with SLP score (61°, 38°, 12°, 8° for Scores 5–2), with the largest decrease at Score 3 indicating that even minor errors reflect reduced differentiation.

Lingual Workspace Maps (LWMs). Per-keypoint KDE density heatmaps with convex hull workspace boundaries (Fig. 3), adapting the convex hull workspace area metric from motor assessment [17] to lingual movement, show hulls expanding and overlapping from Score 5 to 1. Each hull demarcates the spatial workspace a keypoint occupies, where expanding area reflects less precise articulatory targeting, while increasing inter-keypoint overlap indicates loss of spatial distinctiveness between tongue regions.

B. Speaker-Independent Classification

Table II and Fig. 4a present classification results measuring agreement with expert SLP scores. Audio achieves LOSO AUC of 0.687, whereas ultrasound alone performs near chance (0.508) due to inter-speaker anatomical variation. Naive fusion *degrades* to 0.661 because anatomy-encoding features confound the model with speaker identity, while selective fusion (audio + 48 kinematic US features) achieves the best AUC of

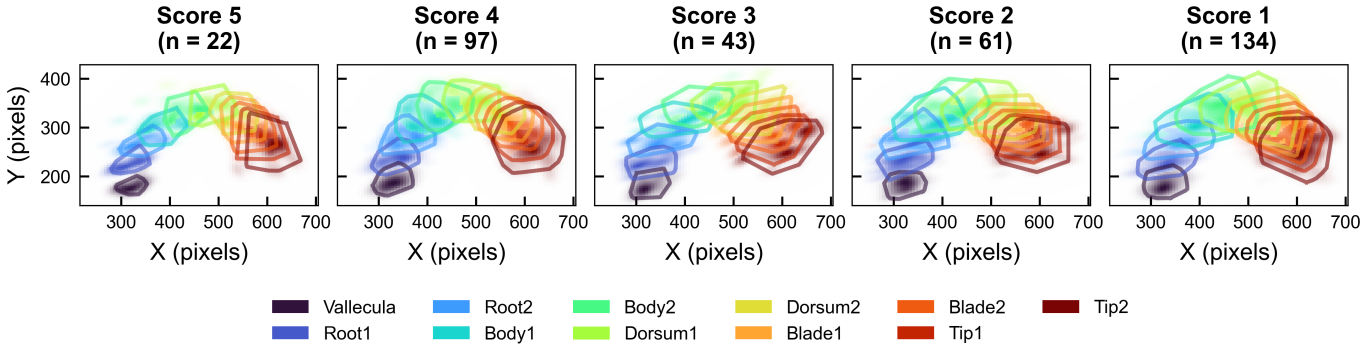


Fig. 3. Lingual workspace maps (LWMs): score-stratified KDE density heatmaps with convex hull workspace boundaries per tongue keypoint, colored by contour position. Hulls expand from Score 5 to 1, indicating loss of spatial precision.

TABLE II
BINARY CLASSIFICATION (SCORE = 5 VS 1–4) AND FEATURE GROUP ABLATION UNDER GKF AND LOSO EVALUATION.

Configuration	Feat.	GKF AUC	LOSO AUC
Ultrasound only	238	0.474	0.508
Audio only	450	0.712	0.687
Naive fusion	688	0.663	0.661
Selective fusion	**498**	**0.736**	**0.725**
Audio + single US group ablation			
+ Anatomical†	490	0.665	0.730
+ Positional†	490	0.691	0.700
+ Shape†	560	0.719	0.724
+ **Kinematic**	**498**	**0.736**	**0.725**
+ All US (naive)	688	0.672	0.672

†Static features; improve LOSO but degrade or stagnate GKF.

0.725 by capturing how the tongue *moves* rather than its static anatomy.

Feature group ablation (Table II, bottom) validates this selection. All four US groups improve LOSO AUC when fused individually with audio, but only kinematic features also improve GKF AUC (0.736 vs 0.712). Anatomical features show the largest GKF–LOSO discrepancy (0.665 vs 0.730), indicating unstable generalization driven by speaker-specific anatomy. Multi-group combinations (e.g., kinematic + shape) do not improve over kinematic alone under GKF, confirming that each additional static group degrades generalization. Combining all groups degrades both protocols, confirming that the 190 static features overpower the kinematic signal.

C. Feature Analysis

SHAP analysis (Fig. 4b) shows that while audio spectral features carry the primary signal, kinematic features (tip/dorsum speed, jerk) add movement information that audio alone cannot capture. By excluding anatomy-encoding features, the model cannot exploit speaker identity as a spurious correlation.

D. Selective Fusion Rescue Analysis

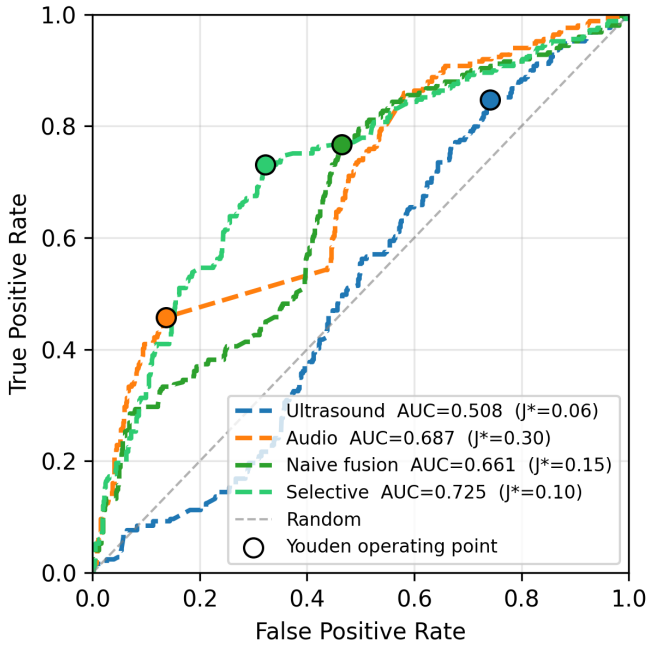
Selective fusion agrees with the SLP judgment on 68 phoneme instances that audio alone misclassifies as impaired, with *zero* lost in the other direction. These rescued instances are dominated by velar stops (/g/: 38, /k/: 29, /r/: 1), where children achieve correct tongue dorsum raising but produce acoustically atypical output—consistent with the covert contrast scenario where the tongue moves correctly but the sound is atypical. The near-absence of /r/ rescue (1 of 68) is expected because rhotic errors involve incorrect tongue positioning where both modalities agree, and because /r/ admits multiple articulatory strategies (bunched, retroflex) not distinguished in the annotations. This asymmetry means fusion allows a model to distinguish "sounds wrong but moves right" (rescuable velars) from "sounds wrong and moves wrong" (true articulatory errors). Per-speaker rescue rates reach ∼40% for three speakers, indicating systematic acoustic atypicality despite correct articulation [10] where ultrasound biofeedback is most needed [2].

V. DISCUSSION AND CONCLUSION

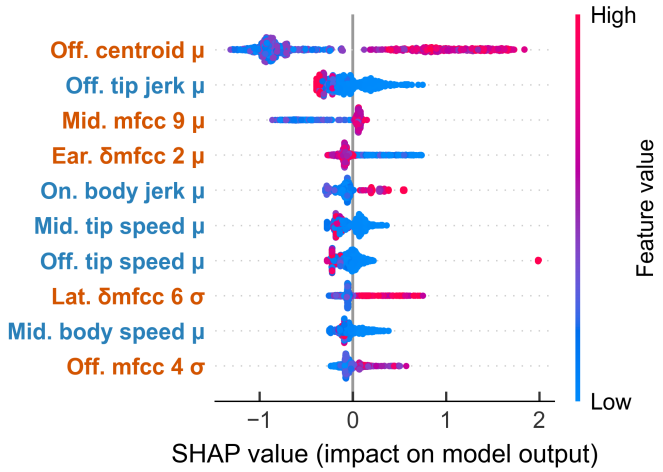
PULSE shows that principled feature selection over interpretable, grouped features improves speaker-independent SSD assessment over naive fusion. Named, grouped features made it possible to identify and exclude the 190 static features encoding speaker identity, which end-to-end approaches cannot do. LDI and LWMs provide new ways to characterize the motor basis of SSD, producing population- and session-level visualizations that can support clinical interpretation.

However, these group-level patterns do not translate directly to speaker-independent classification with only 9 speakers, owing to the 190 static features that overwhelm 48 kinematic features and the classifier over-relies on speaker-identifying anatomy. Selective fusion resolves this by retaining only movement dynamics that generalize across speakers. The 68 rescued velar instances, correct articulations that sound atypical, provide complementary information (68 gained, 0 lost). In biofeedback therapy, where the ultrasound probe is already present, this signal requires no additional instrumentation and is relevant to the covert contrast problem [10]: for three of the nine speakers, ∼40% of correct productions would be missed by audio-only screening.

Limitations. With only 9 SSD speakers, statistical power is limited and results may not generalize to larger or more



(a) LOSO ROC curves with per-configuration Youden J^* operating points marked. Selective fusion achieves the highest AUC (0.725); naive fusion (0.661) performs worse than audio alone (0.687).



(b) SHAP beeswarm (top 10 features, selective fusion). Orange = audio, blue = US kinematic. Kinematic features appear alongside spectral features, confirming complementarity.

Fig. 4. Classification analysis.

diverse populations. The binary threshold (score = 5 vs 1–4) discards severity grading, and predicting the full 5-level scale is a direction for future work. The DLC model was pretrained on adult speakers. Child-specific fine-tuning could improve pose accuracy but requires new keypoint annotations. Features are segment-level aggregates. Sequence models operating on full articulatory trajectories could capture temporal dynamics directly, also enabling exploration of earlier, richer fusion schemes. Finally, validation on additional corpora is needed to confirm cross-dataset generalizability. Future work includes broader baselines and fusion architectures, and a clinical user study with speech-language pathologists to establish real-world utility.

References

[1] Y. Wren, L. L. Miller, T. J. Peters, A. Emond, and S. Roulstone, "Prevalence and predictors of persistent speech sound disorder at eight years old: Findings from a population cohort study," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 4, pp. 647–673, 2016.

[2] J. Cleland, J. M. Scobbie, and A. A. Wrench, "Using ultrasound visual biofeedback to treat persistent primary speech sound disorders," *Clinical Linguistics & Phonetics*, vol. 29, no. 8-10, pp. 575–597, 2015.

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

[4] A. Wrench and J. Balch-Tomes, "Beyond the Edge: Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut," *Sensors*, vol. 22, no. 3, p. 1133, Jan. 2022.

[5] H. Kabakoff, D. Harel, M. Tiede, D. H. Whalen, and T. McAllister, "Extending Ultrasound Tongue Shape Complexity Measures to Speech Development and Disorders," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 7, pp. 2557–2574, Jul. 2021.

[6] M. S. Ribeiro, J. Cleland, A. Eshky, K. Richmond, and S. Renals, "Exploiting ultrasound tongue imaging for the automatic detection of speech articulation errors," *Speech Communication*, vol. 128, pp. 24–34, Apr. 2021.

[7] X. Dan, K. Xu, Y. Zhou, C. Yang, Y. Chen, Y. Dou, and C. Yang, "Spatio-temporal masked autoencoder-based phonetic segments classification from ultrasound," *Speech Communication*, vol. 169, p. 103186, 2025.

[8] S. Al Ani, J. Cleland, and A. Zoha, "Deep learning in ultrasound tongue imaging: a systematic review toward automated detection of speech sound disorders," *Frontiers in Artificial Intelligence*, vol. 8, p. 1631134, 2025.

[9] I. Aytutuldu, Y. Genc, and Y. S. Akgul, "Audio-only phonetic segment classification using embeddings learned from audio and ultrasound tongue imaging data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4501–4515, 2024.

[10] J. Cleland, J. M. Scobbie, C. Heyde, Z. Roxburgh, and A. A. Wrench, "Covert contrast and covert errors in persistent velar fronting," *Clinical Linguistics & Phonetics*, vol. 31, no. 1, pp. 35–55, Jan. 2017.

[11] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning," *Nature Neuroscience*, vol. 21, no. 9, pp. 1281–1289, 2018.

[12] N. E. Helwig, S. Hong, E. T. Hsiao-Wecksler, and J. D. Polk, "Methods to temporally align gait cycle data," *Journal of Biomechanics*, vol. 44, no. 3, pp. 561–566, 2011.

[13] K. M. Dawson, M. K. Tiede, and D. H. Whalen, "Methods for quantifying tongue shape and complexity using ultrasound imaging," *Clinical Linguistics & Phonetics*, vol. 30, no. 3-5, pp. 328–344, 2016.

[14] S. Balasubramanian, A. Melendez-Calderon, A. Roby-Brami, and E. Burdet, "On the analysis of movement smoothness," *Journal of NeuroEngineering and Rehabilitation*, vol. 12, p. 112, 2015.

[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[17] K. Tyryshkin, A. M. Coderre, J. I. Glasgow, T. M. Herter, S. D. Bagg, S. P. Dukelow, and S. H. Scott, "A robotic object hitting task to quantify sensorimotor impairments in participants with stroke," *Journal of NeuroEngineering and Rehabilitation*, vol. 11, p. 47, 2014.