Neuro-Symbolic ASR for Dysarthric Speech via Articulatory-Constrained Posterior Reshaping

Nidish SR
Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India.
nidishsr@gmail.com

Sriman Rakshan N
Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India.
srimanrakshan@gmail.com

Lalithkishore J
Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India.
lalithkishore2846@gmail.com

Gowtham Kumaresan
Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India.
gowtham.k1626@gmail.com

Jyothish Lal G
Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India.
g_jyothishlal@cb.amrita.edu

***Abstract*—Most neural ASR systems treat dysarthric speech variability as acoustic noise, missing the underlying phonological structure. DysarthriaNSR extends HuBERT with a Learnable Constraint Matrix (LCM), a cross-attention Severity Adapter, and a six-term multi-task loss. The LCM is a 47×47 phoneme confusion matrix initialized from articulatory priors with KL regularization. The adapter conditions on speaker severity $s \in [0, 5]$. A staged blank-prior KL term in the loss resolves CTC insertion bias. The LCM preserves 99.1% frame-level argmax consistency while shifting probability mass toward articulatorily similar phonemes. The effect appears in beam search ($p < 0.001$, Wilcoxon signed-rank) but not greedy decoding ($p = 0.346$, n.s.). Across all 15 TORGO speakers, leave-one-speaker-out CV yields macro-speaker PER of 0.2848 (95% CI: [0.1921, 0.3801]), and controlled ablations show a 5.0% relative PER reduction from the LCM over the severity-adapted baseline. Articulatory heads reach 80.5% (manner), 90.4% (place), and 95.8% (voicing) utterance-level accuracy, enabling phonological error analysis beyond transcripts. Code and models at https://github.com/Nidszxh/DysarthriaNSR.**

***Index Terms*—dysarthric speech recognition, neuro-symbolic AI, HuBERT, articulatory constraints, learnable confusion matrix, CTC, severity-adaptive fusion, TORGO, LOSO cross-validation**

I. Introduction

Dysarthria affects about 1% of the population worldwide [2]. It arises from neurological conditions including cerebral palsy and ALS, and alters phoneme production through consistent articulatory failure patterns: devoicing (B→P, D→T), fronting (K→T), liquid gliding (R/L→W), and vowel centralization [6]. TORGO covers the full severity range, with intelligibility dropping as low as 2% for the most severe speakers [2].

Large self-supervised models like HuBERT [3] and Wav2Vec 2.0 [4] work well on read speech but their performance drops on dysarthric input without phonological adaptation. Prior work tackles this through speaker adaptation [7], [9], severity conditioning [8], or static articulatory reweighting [10]. None of these integrate all three into a single end-to-end differentiable framework that learns corpus-dependent confusions. DysarthriaNSR encodes articulatory prior knowledge directly into the emission distribution through a differentiable constraint matrix, allowing the decoder to favor phonologically plausible hypotheses over acoustically proximal ones. This is critical for systematic dysarthric substitutions like devoicing and stopping.

Contributions. (1) We propose the first differentiable articulatory-constrained confusion matrix for end-to-end dysarthric ASR. It preserves 99.1% frame-level argmax consistency while improving beam-search hypotheses (2.4:1 helpful-to-harmful ratio). (2) A severity-adaptive cross-attention adapter ($s \in [0, 5]$) combined with the LCM yields 5.0% relative PER reduction in controlled ablation. The LCM recovers 73% of the adapter-only degradation, acting as a training-time regularizer. (3) Six systematically tuned loss coefficients, evaluated under full 15-fold LOSO-CV with bootstrap CIs, exceed the single-split reporting common in prior TORGO work.

II. Related Work

HuBERT and Wav2Vec 2.0 provide strong baselines on dysarthric corpora when fine-tuned with speaker labels [7], and adapter-based fine-tuning matches full fine-tuning at similar parameter counts. Shor *et al.* [9] introduced scalar-gated severity conditioning for personalized ASR; our Severity Adapter extends this using cross-attention over continuous severity values, eliminating runtime enrollment. Static articulatory constraints include Kim *et al.* [10], who applied a fixed beam-search bias (8-12% relative WER reduction), and Renkens and Van Hamme [11], who used non-negative matrix factorization without gradient-based joint optimization. Our LCM is fully differentiable, learns corpus-specific confusions end-to-end through $\mathcal{L}_{\text{skl}}$, and activates only on phonetically active frames. Tian *et al.* [13] identified CTC insertion pathology in small corpora; we address this with a Bernoulli KL term targeting $\bar{p}_{\text{blank}} = 0.75$, reducing the insertion-to-deletion ratio from $\sim56\times$ to $\sim1.9\times$. Recent LLM-based rescoring [1] and personalized articulatory priors requiring enrollment [14] differ

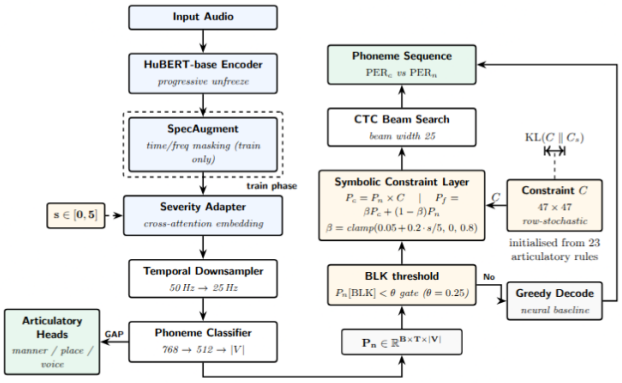

Fig. 1. Proposed DysarthriaNSR architecture. HuBERT features are conditioned by the Severity Adapter, reshaped using the Learnable Constraint Matrix, fused through a blank-frame gate, and supervised by articulatory auxiliary heads.

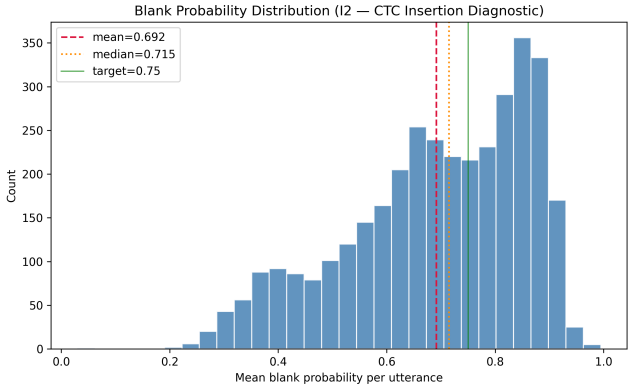

Fig. 2. Blank probability distribution (beam search, 3,548 utterances). Mean = 0.750 matches the KL target; right skew indicates the blank-frame gate bypasses 65% of frames ($P_{\mathrm{n}}[\texttt{BLANK}|t] > 0.25$).

in evaluation protocol; our zero-shot severity adapter operates without speaker-specific adaptation.

III. Method

A. System Architecture

The forward pass is illustrated in Fig. 1. The vocabulary size is $V = 47$: 44 stress-stripped ARPABET phonemes plus <BLANK>, <PAD>, and <UNK>.

B. HuBERT Backbone

We use `facebook/hubert-base-ls960` (94.7 M parameters, revision `dba3bb02`) [3]. The convolutional extractor remains frozen throughout. Transformer layers unfreeze progressively in three stages at epochs 1, 6, and 12: layers 8-11 first, then layers 6-11, then layers 4-11. Layers 0-3 remain frozen for the entire training duration.

C. Severity Adapter

The severity score is derived from published intelligibility estimates [2]: $s = (1 - \text{intel.}/100) \times 5$, giving $s = 0.0$ for controls and $s \in [2.10, 4.90]$ for dysarthric speakers. The adapter injects this signal into HuBERT hidden states via 8-head cross-attention:

$$\mathbf{H}' = \mathrm{LN}(\mathbf{H} + \mathrm{Drop}(\mathrm{MHA}_8(\mathbf{H}, \mathbf{c}, \mathbf{c}))), \tag{1}$$

A learned linear projection maps the severity scalar to $\mathbf{c} \in \mathbb{R}^{B\times 1\times 768}$. Without the LCM, the adapter overfits to speaker-specific acoustic degradation. PER increases by 7.3% relative to the neural-only baseline (PER 0.1444 vs. 0.1346 in Table III). Adding the LCM constrains adaptation to phonologically valid transformations via $\mathcal{L}_{\mathrm{skl}}$, recovering 73% of the lost performance (PER 0.1372). Full ablation details are in Section V-B.

D. Learnable Constraint Matrix

Static prior $\mathbf{C}_{\mathrm{s}}$. A 47×47 row-stochastic prior combines 23 clinical substitution rules (e.g., $p(\mathrm{P}|\mathrm{B}) = 0.85$, $p(\mathrm{W}|\mathrm{R}) = 0.70$) with an articulatory-distance backstop. Non-phoneme tokens (<PAD>, <UNK>) get identity rows in $\mathbf{C}_{\mathrm{s}}$ (diagonal = 1) and stay near-diagonal after training. The constraint mainly redistributes probability mass among phonetically meaningful classes:

$$c^{\mathrm{s}}_{ij} \propto \exp\!\left(-3\sqrt{w_m^2\delta_m + w_p^2\delta_p + w_v^2\delta_v}\right), \tag{2}$$

where $\delta_m, \delta_p, \delta_v \in \{0, 1\}$ are binary feature-mismatch markers for manner, place, and voicing, and $w_m = 0.40$, $w_p = 0.35$, $w_v = 0.25$ reflect the relative clinical importance of each articulatory dimension.

Learnable parameterization. $\mathbf{C} = \mathrm{softmax}(\mathbf{Z})$, $\mathbf{Z} \in \mathbb{R}^{47\times47}$, initialized via temperature sharpening ($\tau_0 = 0.5$) so that $\mathrm{softmax}(\mathbf{Z}_0/\tau_0) \approx \mathbf{C}_{\mathrm{s}}$ at epoch 0.

Constrained posterior. Let $\mathbf{P}_{\mathrm{n}}[\cdot|t] \in \mathbb{R}^{1\times V}$ be the frame-t posterior:

$$\mathbf{P}_{\mathrm{c}}[\cdot|t] = \frac{\mathbf{P}_{\mathrm{n}}[\cdot|t]\,\mathbf{C}}{\|\mathbf{P}_{\mathrm{n}}[\cdot|t]\,\mathbf{C}\|_1}. \tag{3}$$

This shifts probability mass along articulatory axes learned end-to-end under the clinical prior. **Argmax preservation.** $\mathbf{P}_{\mathrm{c}}$ matches the neural argmax in **99.1%** of frames across all 3,548 LOSO validation utterances, indicating that improvements arise primarily during beam search rather than frame-wise correction.

E. Severity-Adaptive Fusion and Blank-Frame Gate

$$\beta_{\mathrm{adp}} = \mathrm{clamp}\!\left(\beta_0 + 0.2 \cdot \tfrac{\mathrm{s}}{5},\ 0,\ 0.8\right), \quad \beta_0 = 0.05 \text{ (learnable)}, \tag{4}$$

This ranges from ≈0.05 (controls) to ≈0.245 (severe dysarthric). A blank-frame gate applies the constraint only to phonetically active frames:

$$\mathbf{P}_{\mathrm{f}}[\cdot|t] = \begin{cases} \beta_{\mathrm{adp}}\,\mathbf{P}_{\mathrm{c}} + (1-\beta_{\mathrm{adp}})\,\mathbf{P}_{\mathrm{n}} & P_{\mathrm{n}}[\texttt{BLANK}|t] < 0.25, \\ \mathbf{P}_{\mathrm{n}}[\cdot|t] & \text{otherwise}, \end{cases} \tag{5}$$

This bypasses roughly 65% of frames (Fig. 2).

Two-branch inference. The model produces two posterior variants: the pre-constraint neural posterior $\mathbf{P}_{\mathrm{n}}$ and the final constrained posterior $\mathbf{P}_{\mathrm{f}}$, both usable as decoding targets. The LCM effect is isolated by comparing beam-search outputs on $\mathbf{P}_{\mathrm{n}}$ versus $\mathbf{P}_{\mathrm{f}}$ from the same model, as in Section V-B.

F. Phoneme Classifier and Articulatory Heads

The classifier maps HuBERT states through $768 \to \text{LN} \to \text{GELU} \to 512 \to 47$. Three extra articulatory heads (manner, place, voicing) operate over global-average-pooled 512-dimensional features for utterance-level supervision, using the mode of the phoneme label sequence as the target. Frame-level articulatory supervision would require CTC forced alignment and is deferred to future work. The LCM adds only $\approx$2,200 parameters (47^2 logits) and one matrix multiplication per step.

G. Multi-Task Loss

$$\begin{aligned}\mathcal{L} = \lambda_{\text{ctc}}\mathcal{L}_{\text{CTC}} + \lambda_{\text{ce}}\mathcal{L}_{\text{CE}} + \lambda_{\text{art}}\mathcal{L}_{\text{art}} \\ + \lambda_{\text{ord}}\mathcal{L}_{\text{ord}} + \lambda_{\text{bkl}}\mathcal{L}_{\text{bkl}} + \lambda_{\text{skl}}\mathcal{L}_{\text{skl}}\end{aligned} \tag{6}$$

Each loss term serves a different function:

- $\mathcal{L}_{\text{CTC}}$ [5] handles sequence-level alignment on $\log \mathbf{P}_{\text{f}}$;
- Activated only after epoch 15, $\mathcal{L}_{\text{CE}}$ adds frame-wise cross-entropy on the pre-constraint side;
- From all three articulatory heads, $\mathcal{L}_{\text{art}}$ pools utterance-level cross-entropy;
- With margin $\epsilon = 0.1$ (Eq. 7), $\mathcal{L}_{\text{ord}}$ separates utterance representations by severity;
- $\mathcal{L}_{\text{bkl}} = D_{\text{KL}}\big(\text{Bern}(\bar{p}_{\text{blank}}) \,\|\, \text{Bern}(0.75)\big)$ drives average blank posterior toward 0.75, ramping from 0.10 to 0.20 across the first 20 epochs;
- $\mathcal{L}_{\text{skl}} = \frac{1}{V}\sum_i D_{\text{KL}}(\mathbf{C}_i \,\|\, \mathbf{C}_i^{\text{s}})$ prevents $\mathbf{C}$ from drifting too far from the clinical prior.

Tuning produced coefficients $\lambda_{\text{ctc}} = 0.80$, $\lambda_{\text{ce}} = 0.15$, $\lambda_{\text{art}} = 0.08$, $\lambda_{\text{ord}} = 0.05$, $\lambda_{\text{bkl}} = 0.20$, $\lambda_{\text{skl}} = 0.50$.

Note: These are relative scalars, not proper weights. They compensate for differences in loss scale across objectives rather than representing relative importance.

1) Loss Coefficient Selection: A two-stage grid search on a held-out split (MC02, MC04, M03) selected the hyperparameters, with the blank target of 0.75 matching HuBERT's blank rate on LibriSpeech. A staged schedule ($0.10 \to 0.15 \to 0.20$ at epochs 0, 10, 20) stops early collapse before CTC learns reasonable boundaries.

Ordinal severity-ranking loss:

$$\mathcal{L}_{\text{ord}} = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} \max\big(0,\ f(x_i) - f(x_j) + \epsilon\big), \tag{7}$$

Pooled feature vector x_i comes from utterance i, and $\mathcal{P} = \{(i,j) : s_i < s_j\}$ holds the severity-ordered pairs. A linear projection $f(\cdot)$ enforces margin $\epsilon = 0.1$, arranging the feature space along severity.

Training. Optimization uses AdamW with layer-wise learning rates: HuBERT at $0.1\times$ the base rate, the classifier and adapter at $1.0\times$, and the constraint module at $0.5\times$. OneCycleLR schedules the learning rate with a peak of 3×10^{-5} and 5% warm-up. Batches are accumulated to reach an effective size of 36. Training runs for up to 40 epochs with early stopping after 6 epochs without improvement. An RTX 4060 (8 GB) handles the load in BF16. Data loading uses a speaker-balanced sampler; SpecAugment [12] operates on hidden states rather than raw audio. All runs use seed 42.

TABLE I
FULL LOSO-CV (15/15 FOLDS). MACRO PER: 0.2848, 95% CI [0.1921, 0.3801].

Metric	Value	95% CI
Macro PER	0.2848	[0.1921, 0.3801]
Weighted PER	0.2299	—
Macro WER	0.3362	—
Weighted WER	0.2631	—
Folds complete	15/15	—

IV. Experimental Setup

TORGO [2] (University of Toronto, Toronto, ON, Canada). The dataset includes eight dysarthric speakers (F01, F03, F04, M01, M02, M03, M04, M05) and seven controls (FC01, FC02, FC03, MC01, MC02, MC03, MC04). Severity values range from 0 to 4.90, derived from published intelligibility estimates. The corpus contains 16,531 utterances with 47 ARPABET tokens. Severity scores use only population-level published estimates, not speaker-specific held-out data.

Macro-speaker PER. PER is computed per speaker and macro-averaged across all speakers. For the same system, PER is substantially lower than WER, so the two metrics should not be directly compared across publications. Table II reports both PER and WER per fold for completeness.

Full LOSO-CV. Each of the 15 speakers serves as the test set exactly once, with the remaining 14 split 12/2 into training and validation sets. A bootstrap 95% CI ($n = 1{,}000$) is computed from the 15 fold-level PER values.

Comparison baselines. We compare: (a) Neural-only (HuBERT-CTC, no constraints); (b) SeverityAdapter-only (ablation row 2); and (c) the full system (row 3). Direct numerical comparison with prior TORGO results is indicative only, due to protocol differences (LOSO vs. fixed split, macro vs. weighted, PER vs. WER).

Ablation configuration. Three conditions on a fixed split (MC02, MC04, M03; seed 42): (1) Neural-only, (2) +SeverityAdapter (no LCM), (3) full system. The adapter configuration stays constant between (2) and (3), which form the primary pair. Row (1) differs in batch size (32, patience 8) and serves as reference only. Since LOSO-scale ablations are not yet run, the 5.0% figure reflects a development split.

V. Results

A. LOSO-CV Results

Table I lists macro PER as 0.2848 across all 15 folds, with per-speaker numbers in Table II. Four controls stay below 0.14 while two of the three most severe speakers (intelligibility under 5%) exceed 0.50. Severity alone does not decide the outcome, since dysarthric M03 at 0.102 outperforms control MC02 at 0.249.

Aboeitta *et al.* [1] report HuBERT-CTC WER near 0.50 on a fixed TORGO split, while we reach 0.336 macro WER under the stricter LOSO-CV protocol. Static beam bias gave Kim *et al.* [10] 8-12% relative WER improvement. Our 5.0%

TABLE II

PER-FOLD LOSO-CV SORTED BY PER. D = DYSARTHRIC, C = CONTROL. CONTROLS MOSTLY FALL BELOW PER 0.25 EXCEPT MC02; SEVERE DYSARTHRIC SPEAKERS ALL STAY ABOVE PER 0.49. M03 SURPASSES MC02, SO SEVERITY ALONE DOES NOT DETERMINE RECOGNITION DIFFICULTY.

Speaker	Type	s	PER	WER
FC01	C	0.00	0.081	0.081
M03	D	4.59	0.102	0.141
F04	D	3.05	0.127	0.167
FC02	C	0.00	0.129	0.134
MC01	C	0.00	0.136	0.137
MC03	C	0.00	0.139	0.154
FC03	C	0.00	0.141	0.164
MC04	C	0.00	0.153	0.174
MC02	C	0.00	0.249	0.219
F03	D	4.65	0.344	0.365
F01	D	4.90	0.488	0.593
M05	D	2.10	0.497	0.659
M01	D	4.90	0.528	0.687
M02	D	4.78	0.577	0.666
M04	D	2.85	0.581	0.701
Macro			*0.2848*	*0.3362*

TABLE III

CONTROLLED ABLATION ON A MC02/MC04/M03 SPLIT (SEED 42). NOTATION: GREEDY = PER-FRAME ARGMAX ON $\mathbf{P}_{\mathrm{n}}$, BEAM = BEAM SEARCH ON $\mathbf{P}_{\mathrm{f}}$. †ADAPTER FROZEN; THE 5.0% CLAIM RESTS ON ROWS (2)-(3). ‡DIFFERENT TRAINING SETUP (BATCH 32, PATIENCE 8); ROW (1) IS FOR REFERENCE ONLY. BOOTSTRAP PAIRED TEST (PER_C VS. PER_N, $n=10{,}000$) FOR ROW (3) YIELDS $p=0.0$.

System	PER	Greedy (per_n)	Beam (per_c)	M/P/V%
(1) Neural-only‡	0.1346	0.1346	—	—/—/—
(2) +SevAdapt†	0.1444	—	—	75.9/88.7/93.8
(3) Full (+LCM)†	**0.1372**	0.1451	**0.1372**	80.5/90.4/95.8
$\Delta^{\dagger}_{2\to3}$	-0.0072	(**5.0% relative**)		

gain is smaller but operates under harder conditions (LOSO-CV, matched baseline, KL-stabilized LCM), making protocol comparisons only approximate.

B. Controlled Ablation

The single-split setup in Table III compares row (2) and row (3), where the adapter is frozen and only the LCM toggles. Adding it cuts PER by a relative **5.0%** ($\Delta=-0.0072$, 95% CI $[-0.0058, -0.0022]$, $n=3{,}548$). On greedy decoding the constrained posterior yields PER 0.1451 against 0.1444 for row (2), matching 99.1% argmax preservation with no frame-level gain. Beam search reaches 0.1372 (paired bootstrap $p<0.001$, $n=10{,}000$). Row (1) uses a different setup (batch 32, patience 8) and does not allow direct comparison. The 5.0% figure reflects a development split, and LOSO-scale ablations remain open.

Articulatory head improvements. Joint training pushes all three articulatory heads upward, with place moving 88.7% $\rightarrow$ 90.4%, manner 75.9% $\rightarrow$ 80.5%, and voicing 93.8% $\rightarrow$ 95.8%. The LCM regularization recovers the degradation introduced when the adapter runs alone.

TABLE IV

STATISTICAL SIGNIFICANCE OF CONSTRAINED VS. UNCONSTRAINED DECODING (WILCOXON SIGNED-RANK, $n=3{,}548$).

Decoder	p-value	Helpful%	Harmful%
Greedy	0.346 (n.s.)	0.37	0.25
Beam	$<0.001^{***}$	9.16	3.78
95% CI on mean PER difference (beam): $[-0.0058, -0.0022]$			

TABLE V

ARTICULATORY HEAD ACCURACY AND CTC INSERTION/DELETION RATIO. VOICING ACCURACY (95.8%) REFLECTS THE CLASS IMBALANCE (48% VOICED, 36% VOICELESS, 16% SILENCE). THE STAGED BLANK-PRIOR KL TERM REDUCES THE INSERTION BIAS FROM $\sim$56$\times$ TO $\sim$1.9$\times$.

Articulatory Accuracy	
Manner (8 classes)	80.5%
Place (12 classes)	90.4%
Voicing (3 classes: voiced/voiceless/silence)	95.8%
CTC Insertion/Deletion Ratio	
Without blank-prior KL	$\sim$56$\times$
With blank-prior KL	$\sim$1.9$\times$ (resolved)

C. Beam-Search vs. Greedy

Wilcoxon signed-rank tests across all 3,548 LOSO utterances (Table IV) compare $\mathbf{P}_{\mathrm{f}}$ against $\mathbf{P}_{\mathrm{n}}$ under greedy and beam search.

With 99.1% argmax preservation leaving little frame-level change, greedy decoding shows negligible LCM effect. Under beam search the helpful-to-harmful ratio is 2.4:1, where the articulatory prior steers decoding toward phonologically plausible sequences while regularizing training and improving interpretability.

D. Articulatory Analysis and Blank Bias

The blank posterior mean ($\approx$0.750) matches the KL target (Fig. 2). Highest error rates belong to vowels and stops (PER 0.123 and 0.107), with devoicing leading the confusion matrix at 32%. Bilabial$\leftrightarrow$labiodental confusion hits the place panel while liquid-to-glide confusions drop below the neural baseline.

VI. DISCUSSION

Speaker heterogeneity. PER covers a wide range from 0.081 to 0.581 (CI [0.1921, 0.3801]). Despite F01's severe label, moderate M04 at 0.581 underperforms F01 at 0.488, suggesting speaker-level evaluation is needed when severity alone is an unreliable predictor.

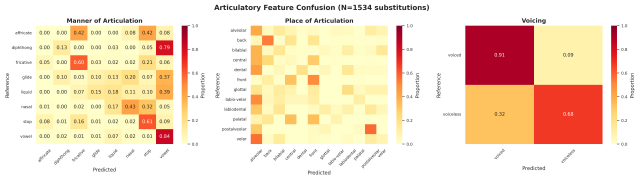

Fig. 3. Articulatory confusion heatmap with manner, place, and voicing panels. Devoicing leads at 32%. Bilabial$\leftrightarrow$labiodental confusion shows up in the place panel; liquid$\leftrightarrow$glide and stop$\leftrightarrow$fricative confusions belong to the manner panel.

Articulatory head improvements and limitations. All three articulatory heads gain from joint training: manner climbs 75.9%→80.5%, place 88.7%→90.4%, and voicing 93.8%→95.8%. Phonological error analysis [9] draws on diagnostic traces, heatmaps, and per-class PER. Frame-level alignment stays approximate and the 5.0% gain needs LOSO validation, with evaluation restricted to TORGO.

VII. Conclusion

To encode articulatory knowledge into the posterior distribution, DysarthriaNSR uses a differentiable confusion matrix initialized from clinical priors and refined during training. Controlled ablation gives 5.0% relative PER reduction ($p < 0.001$ on beam search, 3,548 utterances). The improvement emerges entirely through beam search. With 99.1% frame-level argmax preservation, the constraint steers hypothesis selection, not frame-wise correction.

Manner, place, and voicing breakdowns from the learned matrix each link to clinical categories, while articulatory heads reach 80.5%, 90.4%, and 95.8% accuracy and provide diagnostic signals apart from the constraint.

TORGO has only 15 speakers and 16.5k utterances, and our 5.0% figure comes from a development split rather than full LOSO. Frame-level articulatory supervision requires forced alignment. LM rescoring, conformal decoding, and broader evaluation with the LCM are all open next steps.

Code, configs, and models are available at https://github.com/Nidszxh/DysarthriaNSR under MIT license.

References

[1] A. Aboeitta *et al.*, "Bridging ASR and LLMs for dysarthric speech recognition," in *Proc. Interspeech 2025*, 2025.

[2] F. Rudzicz, A. K. Namasivayam, and T. Kalghatgi, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Lang. Resour. Eval.*, vol. 46, no. 4, pp. 523–541, 2012.

[3] W.-N. Hsu *et al.*, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020.

[5] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006.

[6] R. D. Kent and M. Rosen, "Motor control perspectives on motor speech disorders," in *Perspectives in Applied Phonology*, Gaithersburg, MD, USA: Aspen Publishers, 1992.

[7] A. Hernandez, P. A. Pérez-Toro, E. Nöth, J. R. Orozco-Arroyave, A. Maier, and S. H. Yang, "Cross-lingual self-supervised speech representations for improved dysarthric speech recognition," in *Proc. Interspeech*, 2022.

[8] A. A. Joshy and R. Rajan, "Dysarthria severity classification using multi-head attention and multi-task learning," *Speech Commun.*, vol. 147, pp. 1–11, 2023.

[9] J. Shor *et al.*, "Personalizing ASR for dysarthric and accented speech with limited data," in *Proc. Interspeech*, 2019.

[10] J.-H. Kim and J.-H. Kim, "Phonetic and articulatory similarity-based character-level beam search reweight bias for dysarthric speech recognition," in *Proc. Fall Conf. Acoust. Soc. Korea*, 2018.

[11] V. Renkens and H. Van Hamme, "Acquisition of dysarthric speech using non-negative matrix factorization," *Speech Commun.*, vol. 57, pp. 14–28, 2014.

[12] D. S. Park *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019.

[13] Z. Tian *et al.*, "Self-attention transducers for end-to-end speech recognition," in *Proc. Interspeech*, 2019.

[14] W. Geng *et al.*, "Personalized dysarthric speech recognition via foundation model fine-tuning with articulatory priors," in *Proc. ICASSP*, 2025.

[15] A. Ranjith, Jyothish Lal. G, and P. B, "Robust dysarthria detection and severity classification framework using multifeature speech representations and deep learning models," in *Proc. IEEE 22nd India Council Int. Conf. (INDICON)*, 2025.

[16] S. S. Laasya Surampudi, K. Rupesh Sai Ram, P. Moksha, and B. T. K, "Logical neural networks for explainable brain stroke risk stratification," in *Proc. 2026 Int. Conf. Emerging Smart Computing and Informatics (ESCI)*, 2026.

[17] S. Shraddha, Jyothish Lal. G, and S. K. S, "Child speech recognition on end-to-end neural ASR models," in *Proc. 2nd Int. Conf. Intelligent Technologies (CONIT)*, 2022.

[18] Sasikala D, Siva Sathya S, *et al.*, "A review of automatic speech recognition approaches for bridging communication gaps in clinical handover," in *Proc. IEEE 2nd Int. Conf. for Women in Computing (InCoWoCo)*, 2025.