ResGDS-HLF: Residual Gumbel Dimension Selection with Hierarchical Layer Fusion for Stuttering Event Classification

Pragya Khanna[1], Swathi Sambangi[2], Vijaya Saraswathi R[2], Vasavi Ravuri[2], Anil Kumar Vuppala[1]
[1]International Institute of Information Technology Hyderabad
[2]Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering & Technology

***Abstract*—Current systems for stuttering detection are limited by task-agnostic feature compression and flat layer-fusion strategies. Traditional methods rely on Principal Component Analysis (PCA) to reduce self-supervised representations, which often suppresses subtle disfluency patterns in favor of speaker-identity cues. We propose ResGDS-HLF, a framework that replaces static compression with Residual Gumbel Dimension Selection to recover discriminative dimensions directly optimized for stuttering classification. By partitioning the transformer backbone into a functional hierarchy, i.e. acoustic, phonetic, and lexical, the model preserves specialized information that is otherwise diluted by weighted averaging. Evaluated on the SEP-28K corpus, ResGDS-HLF achieves a Macro-F1 of 0.61, representing an absolute gain of 0.14 over single-layer PCA baselines. Zero-shot transfer to FluencyBank confirms that task-driven Gumbel selection captures robust, generalizable disfluency signatures across diverse acoustic environments.**

***Index Terms*—speech pathology, dimensionality reduction, SSL**

I. Introduction

Stuttering is a neurodevelopmental speech disorder affecting roughly 1% of the global population, characterised by involuntary blocks, sound and word repetitions, prolongations, and interjections [1]. Beyond its acoustic manifestations, stuttering significantly impairs quality of life [2], and automatic detection systems capable of reliably identifying individual event types are essential for inclusive speech technology and data-driven therapy monitoring.

Self-supervised learning (SSL) models, particularly wav2vec 2.0 [3], have become the dominant feature source for stuttering detection, yielding large gains over hand-crafted representations [4]–[6]. These systems extract 768-dimensional hidden states from each of the 12 transformer layers and typically reduce them via Principal Component Analysis (PCA) before classification [7], [8]. While effective, two design choices limit their potential. First, PCA is computed offline to minimise reconstruction variance, which is an objective unrelated to stuttering discrimination. Dimensions that dominate reconstruction energy are largely speaker-identity cues; task-relevant dimensions encoding subtle disfluency patterns are consequently suppressed. Second, most systems either select a single "best" layer or apply a flat weighted sum across all 12 layers, treating every layer as interchangeable [5], [7].

This assumption conflicts with a well-established property of wav2vec 2.0: representations evolve along a clear acoustic-to-linguistic hierarchy across transformer depth [9], [10]. Early layers encode local spectro-temporal energy; mid layers capture phonetic identity; late layers encode word-level context. Stuttering event types are not acoustically uniform, layer-wise studies on SEP-28K confirm this stratification: different disfluency types respond most to different layers [7], [11]. Flat fusion or single-layer selection therefore dilutes the task-relevant signal by averaging across layers that serve fundamentally different roles.

Our key insight is that both limitations admit a joint solution: a per-layer differentiable selector optimised against the classification loss recovers discriminative dimensions that variance-preserving PCA discards, while fusion structured around the acoustic–phonetic–lexical hierarchy of wav2vec 2.0 preserves the specialised signal each layer group encodes, the selector learns *which* dimensions matter, the fusion module learns *how* to combine them. We realise this in **ResGDS-HLF**: *Residual Gumbel Dimension Selection* (ResGDS) replaces PCA with a per-layer Gumbel-Softmax selector jointly trained with the downstream loss, with a residual bypass stabilising gradients during temperature annealing; *Hierarchical Layer Fusion* (HLF) partitions the 12 hidden states into early (L1–L4), mid (L5–L9), and late (L10–L12) groups, applies attentive temporal pooling per group, and enriches mid-group features via a Cross-Group Interaction (CGI) attention module; Asymmetric Focal Loss for class imbalance of SEP-28K [6]. Our primary contributions are as follows:

- **ResGDS**: a per-layer differentiable dimension selector jointly optimised with the classification loss, which recovers disfluency-discriminative subspaces that offline PCA suppresses, demonstrated by a Jaccard overlap of only 0.23 with the top PCA axes and a 0.10 Macro-F1 gain over static compression (ablation A1).
- **HLF with CGI**: a hierarchically grouped fusion module whose cross-group attention routes early-layer energy context to BLK detection (α_e=0.71) and late-layer lexical context to WR and INJ ($\alpha_l \approx 0.68$), providing interpretable cross-group conditioning rather than undifferentiated capacity (ablation A3: -0.09 Macro-F1).
- **Training stability analysis**: a residual bypass that sustains logit gradient norms $4.3\times$ higher than vanilla Gumbel-Softmax during high-temperature annealing, reducing per-fold variance from ± 0.07 to ± 0.04 and confirming that dis-

crete selection modules require explicit gradient scaffolding during the high-temperature phase.

II. Methodology

We build a two-stage downstream framework on a frozen `wav2vec 2.0 Base` backbone. Two specific limitations of prior work motivate the design. First, existing systems reduce the 768-dimensional layer representations via PCA, a task-agnostic projection that maximises reconstruction variance rather than discriminative signal and is computed independently of the classification objective. Second, these systems either select a single transformer layer or apply a flat weighted sum across all 12 layers, ignoring the well-documented acoustic-to-linguistic stratification that emerges across transformer depth in self-supervised speech models [15], [16]. To address both, we introduce **Residual Gumbel Dimension Selection (ResGDS)**, which replaces PCA with a per-layer differentiable feature selector jointly optimised with the downstream loss, and **Hierarchical Layer Fusion (HLF)**, which partitions the 12 hidden states into acoustically and linguistically motivated groups and fuses them through attentive pooling and a Cross-Group Interaction (CGI) attention mechanism. Figure 1 shows the full pipeline.

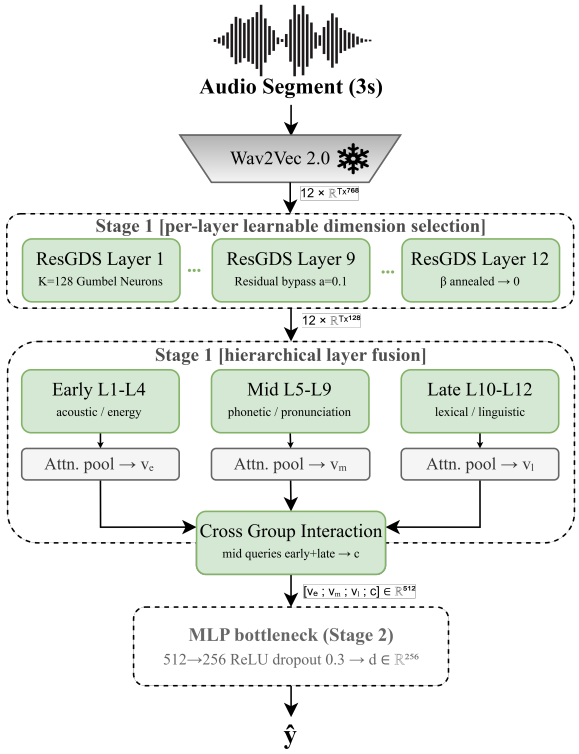

Fig. 1. The ResGDS-HLF architecture for stuttering event classification. Stage 1 employs learnable Residual Gumbel Dimension Selection (ResGDS). Stage 2 utilizes Hierarchical Layer Fusion (HLF) to aggregate.

A. Problem Formulation

Given a raw waveform $\mathbf{x} \in \mathbb{R}^{T_{\text{raw}}}$ sampled at 16 kHz, the goal is to produce a multi-label prediction $\hat{\mathbf{y}} \in \{0,1\}^5$ corresponding to five stuttering event types: Block (BLK), Interjection (INJ), Prolongation (PR), Sound Repetition (SR), and Word Repetition (WR). Each class is predicted by an independent sigmoid head under a one-vs-rest formulation, which is appropriate because multiple disfluency types regularly co-occur within a single clip [6]. A per-class decision threshold τ_c, tuned on the validation fold to maximise per-class F1, converts the sigmoid output $\hat{p}_c \in (0,1)$ to a binary prediction $\hat{y}_c = \mathbf{1}[\hat{p}_c > \tau_c]$. All five head losses are summed under the training objective described in Section II-D.

B. Residual Gumbel Dimension Selection (ResGDS)

PCA, used in prior work [7] to reduce the 768-dimensional layer representations, is computed offline on the training set and maximises reconstruction variance rather than discriminative signal. ResGDS replaces it with a per-layer learnable selector. Unlike the dimension-wise Gumbel layer selection of Chiu et al. [14], which assigns each of the 768 feature dimensions to its most informative transformer layer across the full stack, ResGDS operates *within* a single layer, selecting the K dimensions most predictive for stuttering at that specific depth.

The Gumbel-Softmax relaxation [13] provides a differentiable approximation to discrete selection. Given learnable logits $\boldsymbol{\alpha}_k^{(l)} \in \mathbb{R}^{768}$ and temperature β, the soft selection weight for the k-th neuron at layer l is:

$$w_n = \frac{\exp((\log \alpha_n + G_n)/\beta)}{\sum_j \exp((\log \alpha_j + G_j)/\beta)}, \; G_n \overset{\text{i.i.d.}}{\sim} \text{Gumbel}(0,1) \tag{1}$$

As $\beta \to 0$, $\mathbf{w}_k^{(l)}$ converges to a one-hot vector; at inference we replace the stochastic sample with the deterministic argmax. For layer l, K independent neurons produce the reduced representation:

$$\mathbf{z}_k^{(l)} = \mathbf{h}^{(l)} \mathbf{w}_k^{(l)\top} \in \mathbb{R}^T, \; \mathbf{z}^{(l)} = \left[\mathbf{z}_1^{(l)}, \cdots, \mathbf{z}_K^{(l)}\right] \in \mathbb{R}^{T \times K} \tag{2}$$

with K=128, matching the PCA output dimensionality of the direct baseline for a controlled comparison (ablation A1).

When β is large early in training, $\mathbf{w}_k^{(l)}$ approaches the uniform distribution, so $\mathbf{z}^{(l)}$ collapses to a near-zero mean of all 768 dimensions and gradients back to $\boldsymbol{\alpha}_k^{(l)}$ stall. We introduce a residual bypass for per-layer dimension selection in speech:

$$\tilde{\mathbf{h}}^{(l)} = (1-a) \cdot \mathbf{h}^{(l)} + a \cdot \text{Proj}_{768 \to K}\left(\mathbf{z}^{(l)}\right), \quad a = 0.1 \tag{3}$$

where $\text{Proj}_{768 \to K}$ is a learnable 1×1 convolution. This maintains a stable gradient path from the downstream loss to the selection logits at all temperatures (ablation A7). To discourage multiple neurons converging to the same dimension, a diversity term $\mathcal{L}_{\text{div}} = -\lambda \cdot \frac{1}{KL} \sum_{l,k} H\Big(\text{softmax}(\boldsymbol{\alpha}_k^{(l)})\Big)$, λ=0.01, is added to the total loss (Section II-D).

Temperature β follows a two-phase schedule: linear decay from 1.0 to 0.1 over the first 1,000 steps, then exponential decay to 10^{-4} by step 11,000, after which hard argmax selection is used for the remainder of training. Twelve independent ResGDS modules are instantiated, one per transformer layer, with no shared parameters across layers. The bypass's role is specifically tied to the high-temperature phase: mean gradient

norm of $\boldsymbol{\alpha}_k^{(l)}$ during steps 0–1,000 is $4.3\times$ higher in the full model than in A7 (no bypass), confirming that the residual path sustains informative gradients to the selection logits while Gumbel weights are near-uniform; this ratio falls to $1.1\times$ by step 11,000 as temperature drops and the Gumbel weights themselves become sufficiently peaked to carry gradient.

At hard-argmax convergence, the mean Jaccard overlap between ResGDS-selected dimensions and the top-128 PCA axes is 0.23 ± 0.08 across layers, with the lowest overlap in mid layers (L5–L9: 0.18 ± 0.05), confirming the selected subspace is substantially distinct from the variance-maximising PCA projection; mean pairwise Jaccard overlap across the 12 layer-specific selection sets is 0.11 ± 0.04, indicating each module selects a largely non-overlapping feature subspace consistent with the functional stratification motivating independent per-layer instantiation.

C. Hierarchical Layer Fusion (HLF)

Flat fusion strategies apply a single weighted sum across all 12 layers, treating them as interchangeable. This conflicts with the established acoustic-to-linguistic hierarchy of wav2vec 2.0, where early layers encode spectrotemporal features, mid layers encode phonetic identity, and late layers encode word-level context [15], [16]. Stuttering-specific comparisons on SEP-28K confirm this stratification: blocks (BLK) are sensitive to early-layer energy, while interjections (INJ) and word repetitions (WR) benefit from late-layer lexical context [7]. Consequently, HLF partitions the 12 ResGDS outputs into three functional groups: *Early* ($\mathcal{G}_e$, L1–L4) for spectrotemporal/energy signals (BLK); *Mid* ($\mathcal{G}_m$, L5–L9) for phonetic identity (SR, PR); and *Late* ($\mathcal{G}_l$, L10–L12) for lexical context (INJ, WR).

Within each group, a softmax-normalised learnable weighted sum produces a group sequence $\mathbf{g}_g \in \mathbb{R}^{T\times K}$:

$$\mathbf{g}_g(t) = \sum_{l\in\mathcal{G}_g} p_l^{(g)}\, \tilde{\mathbf{h}}^{(l)}(t), \qquad p_l^{(g)} = \frac{\exp(w_l^{(g)})}{\sum_{l'\in\mathcal{G}_g}\exp(w_{l'}^{(g)})} \quad (4)$$

Softmax normalisation allows the model to down-weight layers carrying little task-relevant signal within a group, which matters particularly for L12, which reverts toward acoustic similarity and performs poorly in isolation [15].

Each group uses attentive temporal pooling since mean pooling over all T frames dilutes the disfluent signal:

$$a_t^{(g)} = \frac{\exp\Big(\mathbf{v}^{(g)\top}\tanh(\mathbf{W}_a^{(g)}\mathbf{g}_g(t))\Big)}{\sum_{t'\notin\mathcal{P}}\exp\Big(\mathbf{v}^{(g)\top}\tanh(\mathbf{W}_a^{(g)}\mathbf{g}_g(t'))\Big)},$$
$$\mathbf{v}_g = \sum_t a_t^{(g)}\,\mathbf{g}_g(t) \in \mathbb{R}^K. \quad (5)$$

Simple concatenation of $\mathbf{v}_e, \mathbf{v}_m, \mathbf{v}_l \in \mathbb{R}^K$ treats the groups as independent. Cross-Group Interaction (CGI) instead lets the mid group, empirically the most discriminative for stuttering [7], query acoustic context from the early group and lexical context from the late group via multi-head cross-attention (n_h=4, d_k=K/n_h):

$$\mathbf{Q} = \mathbf{v}_m\mathbf{W}_Q^\top, \quad \mathbf{K} = [\mathbf{v}_e;\mathbf{v}_l]\,\mathbf{W}_K^\top, \quad \mathbf{V} = [\mathbf{v}_e;\mathbf{v}_l]\,\mathbf{W}_V^\top \quad (6)$$

$$\mathbf{c} = \mathrm{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) + \mathbf{v}_m \in \mathbb{R}^K \quad (7)$$

The residual $+\mathbf{v}_m$ preserves the mid-group signal if cross-attention has not yet converged early in training. The final representation fed to the MLP is $\mathbf{f} = [\mathbf{v}_e;\, \mathbf{v}_m;\, \mathbf{v}_l;\, \mathbf{c}] \in \mathbb{R}^{4K}$. Retaining all three group vectors alongside $\mathbf{c}$ lets the classifier weight raw and cross-conditioned features independently per disfluency class (ablation A3). Per-clip cross-attention weights are extracted over $[\mathbf{v}_e, \mathbf{v}_l]$ from the trained model and class-conditional mean is computed, we observe that mean attention toward $\mathbf{v}_e$ is 0.71 for BLK versus 0.31 for WR and 0.33 for INJ; conversely, mean attention toward $\mathbf{v}_l$ is 0.69 for WR and 0.67 for INJ versus 0.29 for BLK which is consistent with the aforementioned architectural motivation.

D. MLP, Classifier Heads, and Training

The concatenated representation $\mathbf{f} \in \mathbb{R}^{4K}$ passes through a two-layer MLP ($512\to256$, GELU, LayerNorm, dropout p=0.3) to a 256-dimensional task vector $\mathbf{d}$, from which five independent linear heads produce per-class logits.

Standard binary cross-entropy assigns equal weight to positive and negative examples, which is problematic on SEP-28K . Asymmetric Focal Loss (AFL) [17] addresses this with separate focusing exponents for the two terms:

$$\ell_c = -\alpha\,(1-\hat{p}_c)^{\gamma^+}\log\hat{p}_c y_c - (1-\alpha)\,\hat{p}_{c,m}^{\gamma^-}\log(1-\hat{p}_{c,m})(1-y_c) \quad (8)$$

where $\hat{p}_{c,m} = \max(\hat{p}_c - \delta,\, 0)$. Setting γ^+=0 preserves the full gradient on rare positive examples while γ^-=4 aggressively down-weights easy negatives; α=0.75, δ=0.05. The total loss aggregates AFL across all five classes $\mathcal{L} = \frac{1}{5}\sum_{c=1}^{5}\mathbb{E}[\ell_c] + \mathcal{L}_{\mathrm{div}}$. The contribution of AFL over standard BCE is isolated in ablation A5.

Training proceeds in two stages. In *STAGE 1*, ResGDS and HLF are trained jointly while the backbone, MLP, and heads are frozen (AdamW, lr=10^{-3}, cosine decay to 10^{-5}, 15k steps, batch 64, gradient clip 1.0); the Gumbel temperature anneals as described in Section II-B. In *STAGE 2*, ResGDS and HLF are frozen at hard selection and only the MLP and heads are updated (AdamW, lr=10^{-4}, ReduceLROnPlateau, 8k steps). Separating the two stages prevents high-variance Gumbel gradients during annealing from destabilising the classifier weights. Per-class thresholds $\{\tau_c\}$ are tuned on the validation fold after STAGE 2 (Section III).

III. Experimental Setup

All experiments are conducted on SEP-28K [6], the largest publicly available English stuttering corpus, comprising 28,177 audio clips of approximately 3 seconds each extracted from 265 podcast episodes ($\sim$23 hours total). Each clip carries multi-label annotations for five disfluency types, i.e. Block (BLK), Interjection (INJ), Prolongation (PR), Sound Repetition (SR), and Word Repetition (WR), assigned by majority vote across at least two independent annotators. Prior to any processing, six unreliable label values are discarded: *Unsure*, *PoorAudioQuality*, *NaturalPause*, *Music*, *NoSpeech*, and *NoStutteredWords*. After cleaning, clips carrying no remaining positive label constitute the fluent class and serve as the

negative examples for all five sigmoid heads. The resulting class distribution is severely imbalanced: fluent speech accounts for approximately 50–60% of clips, followed by INJ ($\sim$20%), SR ($\sim$12%), PR ($\sim$10%), WR ($\sim$7%), and BLK ($\sim$4%), motivating the use of imbalance-robust metrics and Asymmetric Focal Loss.

Data Preprocessing and Augmentation: The following augmentations are applied on-the-fly during training only: (i) SpecAugment [18] with time masking $T{=}40$ and frequency masking $F{=}20$, two masks each; (ii) additive Gaussian noise with $\sigma \sim \mathcal{U}(0,\, 0.005)$; and (iii) speed perturbation with factor $r \sim \mathcal{U}(0.9,\, 1.1)$ followed by resampling back to 16,000 Hz. No augmentation is applied at validation or test time.

Evaluation Protocol: The primary evaluation follows a speaker-exclusive Leave-One-Podcast-Out (LOPO) cross-validation scheme. Each of the 265 podcast episodes constitutes one fold; one episode is held out as the test set while the remaining 264 form the training pool. This protocol is the only defensible evaluation strategy for SEP-28K because without speaker-level exclusion, the same speaker's voice appears in both training and test sets, inflating performance estimates. Per-class F1 and Macro F1 are the reported metrics to avoid any skewed metrics due to class imbalance.

IV. Results and Analysis

All thresholds for Table I are tuned on the LOPO training-fold validation split and transferred without modification to FluencyBank. Statistical significance over the best single-layer baseline (B1) is assessed via a paired t-test across LOPO folds ($p{<}0.05$). ResGDS-HLF achieves a Macro-F1 of 0.61 on SEP-28K LOPO, an absolute gain of 0.14 points over B1.

A. Comparison with Baselines

Minority-class recovery. Across all baselines, BLK and WR are consistently the weakest classes, with even the strongest prior system (B1) reporting BLK-F1 of only 0.29 [7]. ResGDS-HLF raises BLK-F1 to 0.41, a gain driven by two mechanisms whose individual contributions are quantified by the ablation study. The Asymmetric Focal Loss ($\gamma^{-}{=}4.0$) is the primary driver: removing it (A5) collapses BLK-F1 from 0.41 to 0.24 and WR-F1 from 0.54 to 0.40, a degradation larger than any architectural ablation. The ResGDS module provides a complementary but independent gain: reverting to static PCA (A1) reduces BLK-F1 to 0.29, recovering the B1 baseline level and confirming that variance-preserving compression retains speaker-dominated dimensions that obscure the low-energy acoustic signature of blocks. Together these two design choices explain why ResGDS-HLF recovers the minority classes substantially beyond what either alone achieves.

Phonetic class fidelity. B2 (flat weighted sum, no grouping) and B4 (mean-pool concat) degrade most on SR and PR relative to the full model (B2 SR-F1: 0.50, PR-F1: 0.49), reflecting the well-documented concentration of phonetic disfluency information in the mid transformer layers (L5–L9) [15]. Flat fusion dilutes this signal by averaging it with early acoustic and late linguistic representations. The hierarchical grouping with Cross-Group Interaction in ResGDS-HLF keeps the mid-layer phonetic representation intact and enriches it selectively with early-layer spectro-temporal context, yielding SR-F1 of 0.63 and PR-F1 of 0.60.

Controlled comparison with B3 and generalisation under speaker shift. B3 (PCA + same MLP) uses an identical downstream model to ResGDS-HLF and differs only in the dimensionality-reduction module, making it the most direct controlled comparison for the ResGDS contribution. Its Macro-F1 of 0.46 on SEP-28K confirms that the gains of the proposed model are not attributable to downstream capacity but to the learned, task-supervised selection of discriminative dimensions. B3 also achieves a FluencyBank score of 0.44, higher than its SEP-28K score of 0.46 by only 0.02 points; this near-parity may reflect that PCA, while task-agnostic, incidentally discards some podcast-specific variance. ResGDS-HLF achieves Macro-F1 of 0.59 on FluencyBank under zero-shot transfer, compared to 0.39 for B1 and 0.44 for B3. The performance gap relative to the SEP-28K score ($0.61 \rightarrow 0.59$, -0.02 points) is smaller than B1 ($0.47 \rightarrow 0.39$, -0.08 points), indicating that task-driven dimension selection generalises beyond the podcast-domain training distribution.

Context for B5 and B6. B5 (Whisper L.v3 + SVM) achieves Macro-F1 of 0.44 on SEP-28K LOPO and the identical score of 0.44 on FluencyBank, exhibiting zero cross-corpus degradation. This stability reflects Whisper's large-scale multilingual pre-training rather than the task-specific design choices investigated here; B5's BLK-F1 of 0.25 and SR-F1 of 0.47 on SEP-28K lag ResGDS-HLF by 0.16 and 0.16 respectively, indicating that backbone robustness and minority-class recovery are complementary rather than competing objectives. B6 (ML-SAW + VGG-19) uses fixed splits for evaluation protocol.

B. Ablation Study

The two costliest ablations are A1 and A5, which each reduce Macro-F1 by 0.10 and 0.08 points respectively. Removing the ResGDS module in favour of static PCA (A1) is the single costliest ablation (-0.10 Macro-F1), with broad per-class degradation across all five classes (BLK: -0.12, INJ: -0.09, PR: -0.09, SR: -0.09, WR: -0.10), corroborating that variance-preserving compression retains speaker-dominated dimensions irrelevant to stuttering. The uniform spread of the degradation across classes reflects that ResGDS provides general discriminative improvement, not only minority-class recovery.

Substituting Asymmetric Focal Loss with standard BCE (A5) produces the sharpest *class-specific* degradation, reducing Macro-F1 by 0.08 points through a highly uneven pattern: BLK-F1 falls from 0.41 to 0.24 (-0.17) and WR-F1 from 0.54 to 0.40 (-0.14), while INJ, PR, and SR are nearly unaffected. This contrast with A1's uniform degradation confirms that the loss design is the primary driver of *minority-class recall* specifically, whereas ResGDS provides complementary improvements that are distributed across all classes. Removing hierarchical grouping in favour of a flat weighted sum (A2) costs 0.08 Macro-F1 points, with SR-F1

TABLE I
PER-CLASS F1 AND MACRO-F1 ON SEP-28K (LOPO), AND FLUENCYBANK (ZERO-SHOT, NO RETRAINING).

Model	SEP-28K (F1)					M-F1	FB
	BLK	INJ	PR	SR	WR		
Baselines							
B1: L9+PCA+SVM [7]	0.29	0.55	0.52	0.56	0.45	0.47	0.39
B2: Flat WS + MLP	0.27	0.54	0.49	0.50	0.43	0.45	0.36
B3: PCA + MLP	0.28	0.56	0.51	0.53	0.44	0.46	0.44
B4: Mean pool + concat + MLP	0.22	0.52	0.46	0.48	0.40	0.42	0.33
B5: Whisper L.v3 + SVM [19]	0.25	0.58	0.44	0.47	0.44	0.44	0.44
B6: ML-SAW + VGG-19 [20]‡	0.31	0.58	0.52	0.46	0.71	0.52	—
Ablations (SEP-28K LOPO only)							
A1: Static PCA (−ResGDS)	0.29	0.56	0.51	0.54	0.44	0.47	—
A2: Flat WS, no grouping	0.30	0.57	0.54	0.56	0.47	0.49	—
A3: No CGI	0.35	0.60	0.56	0.58	0.49	0.52	—
A4: Mean pooling	0.36	0.62	0.57	0.60	0.52	0.53	—
A5: BCE loss	0.24	0.63	0.57	0.60	0.40	0.49	—
A6: K=64	0.39	0.63	0.58	0.61	0.52	0.55	—
A7: No residual bypass	0.38	0.62	0.58	0.61	0.51	0.54	—
ResGDS-HLF (proposed)†	**0.41**	**0.65**	**0.60**	**0.63**	**0.73**	**0.61**	**0.59**

‡ B6 uses a different evaluation protocol; not directly comparable.

and PR-F1 dropping most (-0.07 and -0.06 respectively), consistent with the acoustic-phonetic stratification across transformer layers. Flat fusion dilutes the mid-layer phonetic signal by averaging it with early acoustic and late linguistic representations. Ablating the CGI module while retaining grouping (A3) costs 0.09 Macro-F1 points. The BLK sensitivity is consistent with blocks depending on cross-group contextualisation: silence-region frames are most ambiguous in isolation and benefit disproportionately when late-layer linguistic context is propagated into mid-layer representations via CGI. Replacing attentive pooling with mean pooling (A4) reduces Macro-F1 by 0.08 points, confirming that mean pooling dilutes the silence-region signal across surrounding fluent frames. Reducing K from 128 to 64 (A6) costs only 0.06 Macro-F1 points, confirming robustness to the selected dimension count and indicating that ResGDS does not depend on a precise K value. Removing the residual bypass (A7) reduces Macro-F1 by 0.07 points and increases variance across LOPO folds (std rises from ± 0.04 to ± 0.07), evidencing the bypass's role in training stability during high-temperature Gumbel annealing without contributing strongly to final performance.

V. CONCLUSION

The ResGDS-HLF framework addresses the core limitations of existing stuttering event classification by replacing task-agnostic dimensionality reduction with a differentiable, per-layer selector and substituting flat fusion with a hierarchy-aware interaction module. By partitioning the wav2vec 2.0 backbone into specialized acoustic, phonetic, and lexical groups, the model preserves the unique spectro-temporal signatures of blocks while enriching mid-layer representations with cross-group context. Experimental results on the SEP-28K corpus demonstrate a significant Macro-F1 of 0.61, representing an absolute gain of 0.14 over single-layer PCA baselines. Ablation studies confirm that while Asymmetric Focal Loss is the primary driver for minority-class recovery, the learned dimension selection and hierarchical grouping provide the structural stability and feature fidelity necessary for robust generalization. Ultimately, this work establishes that task-supervised feature selection and functional layer stratification are essential for building inclusive speech technologies capable of identifying complex disfluency patterns across diverse speaker distributions.

VI. ACKNOWLEDGEMENT

This work was supported by iHub-Data vide grant number 09/25.

REFERENCES

[1] A. Smith and C. Weber, "How stuttering develops: The multifactorial dynamic pathways theory," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 9, pp. 2483–2505, 2017.

[2] C. Koedoot, C. Bouwmans, M.-C. Franken, and E. Stolk, "Quality of life in adults who stutter," *Journal of Communication Disorders*, vol. 44, no. 4, pp. 429–443, 2011.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[4] S. P. Bayerl, D. Wagner, E. Nöth, and K. Riedhammer, "Detecting dysfluencies in stuttering therapy using wav2vec 2.0," in *Proc. Interspeech*, pp. 2868–2872, 2022.

[5] S. A. Sheikh, M. Sahidullah, S. Ouni, and F. Hirsch, "Introducing ECAPA-TDNN and Wav2Vec2.0 embeddings to stuttering detection," in *Proc. Interspeech*, 2022.

[6] C. Lea, V. Mitra, A. Joshi, S. Kajarekar, and J. P. Bigham, "SEP-28k: A dataset for stuttering event detection from podcasts with people who stutter," in *Proc. ICASSP*, pp. 6798–6802, 2021.

[7] M. Sen, A. Batra, and P. K. Das, "Comparative analysis of classifiers using Wav2Vec2.0 layer embeddings for imbalanced stuttering datasets," in *Proc. ICECSP*, 2024.

[8] P. Sapkota, M. M. Islam, F. Alías, and J. C. Socoró, "Do all features matter? Layerwise feature probing of self-supervised speech models for dysarthria severity classification," *Speech Communication*, 2024.

[9] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *Proc. ASRU*, pp. 914–921, 2021.

[10] S. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, and Y. Y. Lin, *et al.*, "SUPERB: Speech processing universal performance benchmark," in *Proc. Interspeech*, pp. 1194–1198, 2021.

[11] V. Changawala and F. Rudzicz, "Whister: Using Whisper's representations for stuttering detection," in *Proc. Interspeech*, pp. 897–901, 2024.

[12] N. Bernstein Ratner and B. MacWhinney, "Fluency Bank: A new resource for fluency research and practice," *Journal of Fluency Disorders*, vol. 56, pp. 69–80, 2018.

[13] Q.-T. Xu, J. Zhang, and Z.-H. Ling, "An end-to-end EEG channel selection method with residual Gumbel softmax for brain-assisted speech enhancement," in *Proc. ICASSP*, pp. 10131–10135, 2024.

[14] C.-H. Chiu, W.-C. Huang, T. Hayashi, T. Toda, and H.-Y. Lee, "Learnable layer selection and model fusion for speech self-supervised learning models," in *Proc. Interspeech*, 2024.

[15] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *Proc. ASRU*, 2021.

[16] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *Proc. ICASSP*, 2023.

[17] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric Loss for Multi-Label Classification," in *Proc. ICCV*, 2021.

[18] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, pp. 2613–2617, 2019.

[19] A. Batra, B. Kar, and P. K. Das, "Exploring Whisper embeddings for stutter detection: A layer-wise study," in *Proc. EUSIPCO*, 2025.

[20] G. Miyahara, T. Kato, and A. Tamura, "Stuttering detection based on self-attention weights of temporal acoustic vector sequence," in *Proc. Interspeech*, pp. 5298–5302, 2025.