TVP-UNet: Threshold Variance Penalty U-Net for Voice Activity Detection in Dysarthric Speech

Aditya Pandey[1], Tanuka Bhattacharjee[2], Madassu Keerthipriya[3], Darshan Chikktimmegowda[3], Dipti Baskar[3], Yamini BK[3], Seena Vengalil[3], Atchayaram Nalini[3], Ravi Yadav[3], Prasanta Kumar Ghosh[2]

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India
[2]Electrical Engineering Department, Indian Institute of Science, Bengaluru, India
[3]National Institute of Mental Health and Neurosciences, Bengaluru, India

Voice Activity Detection (VAD) & Dysarthria

- **VAD** identifies speech/non-speech regions — critical for ASR, speaker diarization, and speech enhancement.
- **Dysarthria** (caused by Amyotrophic Lateral Sclerosis (ALS), Parkinson's disease (PD)) degrades articulation, phonation, prosody, and speaking rate.
- **Conventional VAD fails** for dysarthric speech:
  - Reduced, unstable intensity makes energy unreliable
  - Irregular articulation blurs speech/silence boundaries
  - A fixed energy/SNR threshold does not generalise across dysarthric speakers
- **Deep learning VAD** also struggles:
  - Requires large amounts of labelled training data
  - Performance degrades under the high acoustic variability of dysarthric speech
- **Implications:** increased false alarms, missed detections, and large performance drop under limited labels.

Our Objective & Contributions

- We propose **TVP-UNet**: a compact U-Net autoencoder with a novel **Threshold Variance Penalty (TVP)** for joint waveform reconstruction and frame-level VAD.
- **TVP loss:** enforces consistency of a statistics-based weak estimator *across multiple thresholds*, stabilising decisions under small waveform variations.
- **Label-efficient:** unified supervised / semi-supervised / unsupervised training via pseudo-labels — near-supervised performance at just 50% labels.
- **First reported VAD** system designed specifically for dysarthric speech (ALS & PD).

Dataset

- **Collection:** NIMHANS, India (ethics-approved; written informed consent)
- **Languages:** 13 Indian languages
- **Task:** Spontaneous speech — a festival they celebrate & a place they recently visited (~1 min each)
- **Annotation:** Speech & silence manually labelled using Audacity
- **Severity:** Rated by 3 Speech-Language Pathologists (SLPs) on 5-point scale; split into **Severe (SV)**, **Mild (ML)**, **No dysarthria (ND)**, **Healthy Controls (HC)**

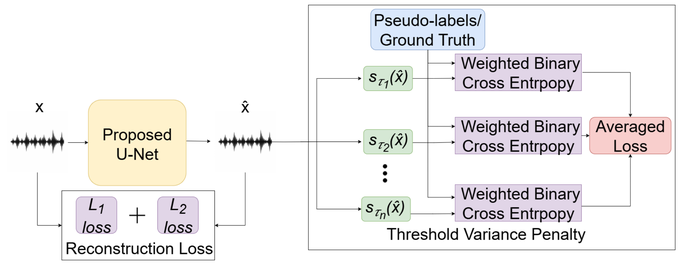
Subject demographics and audio statistics

Group	#Subjects	Age Mean(SD)	Speech seg.(s)	Silence seg.(s)
ALS	230	54.5 (11.56)	1.64 (1.10)	0.83 (1.08)
PD	142	57.1 (9.23)	1.73 (1.06)	1.04 (1.58)
HC	137	43.3 (8.69)	2.47 (2.71)	0.62 (1.02)

Experimental Setup

- **Validation:** 5-fold CV, speaker-independent splits
- **Label regimes:** 0%, 25%, 50%, 75%, 100%
- **Baselines:**
  - Mihalache et al. [1] (supervised DNN)
  - Sarkar et al. [2] (unsupervised, zero-frequency filtering)

Proposed Method: TVP-UNet



x: input speech, $\hat{x}$: reconstructed speech, s_{τ_i}: weak estimator at threshold τ_i

- **Input:** raw 100 ms waveform frames
- **Backbone:** 1D U-Net autoencoder
- **Encoder** (U-Net path): dropout (p=0.3), 3 max-pool downsampling blocks
- **Bottleneck:** layer norm, Gaussian noise, tanh/sigmoid gate (SiLU)
- **Jointly learns** reconstruction and frame-level VAD

Learning Objective:

$$\mathcal{L}_{\text{total}} = \alpha\,\mathcal{L}_{\text{rec}} + (1-\alpha)\,\mathcal{L}_{\text{tvp}}$$

($\alpha \in [0,1]$: trade-off parameter; α=1 ⇒ pure reconstruction, α=0 ⇒ pure TVP; $\alpha_{\min}$: lower bound on α during decay)

- $\mathcal{L}_{\text{rec}}$: reconstruction loss = $\|x-\hat{x}\|_1 + \|x-\hat{x}\|_2^2$; $\mathcal{L}_{\text{tvp}}$: Threshold Variance Penalty
- α decays from 1.0 after midpoint with decay rate ρ=0.9; clipped at $\alpha_{\min}$=0.4

Threshold Variance Penalty (TVP):

Weak estimator (hard decision at threshold τ_k): $\tilde{y}(\hat{x},\tau_k) = \begin{cases} 0, & \frac{|\hat{x}-\mu_{\hat{x}}|}{\sigma_{\hat{x}}} \le \tau_k \\ 1, & \text{otherwise} \end{cases}$

($\mu_{\hat{x}}$: mean of $\hat{x}$; $\sigma_{\hat{x}}$: std. dev. of $\hat{x}$; τ_k: k-th decision threshold; $\tilde{y}$: binary speech/silence label)

Soft differentiable estimator ($\sigma(\cdot)$: sigmoid): $s_{\tau_k}(\hat{x}) = \sigma\left(\frac{|\hat{x}-\mu_{\hat{x}}|}{\sigma_{\hat{x}}} - \tau_k\right)$

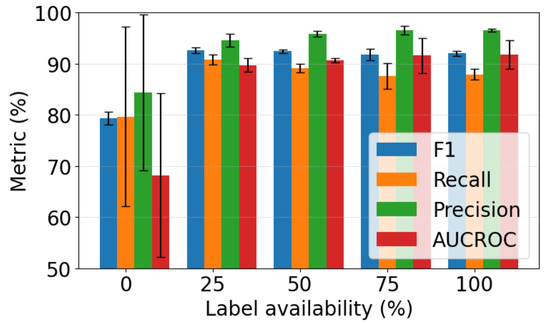
- $\mathcal{T}$={0.05, 0.10, 0.15}: set of K=3 thresholds; $\bar{\tau}$=0.10: mean threshold for pseudo-labels $\tilde{y}(\hat{x},\bar{\tau})$
- L_{WBCE}: weighted binary cross-entropy (handles class imbalance); y: ground-truth label
- **Supervised TVP** (labelled frames): $\mathcal{L}_{\text{tvp}}^{\text{sup}} = \frac{1}{K}\sum_k L_{\text{WBCE}}\left(s_{\tau_k}(\hat{x}), y\right)$
- **Unsupervised TVP** (unlabelled): $\mathcal{L}_{\text{tvp}}^{\text{unsup}} = \frac{1}{K}\sum_k L_{\text{WBCE}}\left(s_{\tau_k}(\hat{x}), \tilde{y}(\hat{x},\bar{\tau})\right)$
- **Per-sample** ($\mathbb{I}_{\text{lab}}$: 1 if labelled, 0 otherwise): $\mathcal{L}_{\text{tvp}} = \mathbb{I}_{\text{lab}}\,\mathcal{L}_{\text{tvp}}^{\text{sup}} + (1-\mathbb{I}_{\text{lab}})\,\mathcal{L}_{\text{tvp}}^{\text{unsup}}$
- Class weight (θ=0.8: smoothing; $N_{\text{pos}}/N_{\text{neg}}$: speech/silence counts; ε: constant): $\beta_t \leftarrow \theta\beta_{t-1} + (1-\theta)\frac{N_{\text{neg}}}{N_{\text{pos}}+\varepsilon}$

References

1. S. Mihalache & D. Burileanu, "VAD with hybrid feature extraction," *Sensors*, vol. 22, 2022.
2. E. Sarkar et al., "Unsupervised VAD via zero-frequency filtering," *Interspeech*, 2022.

Acknowledgement: We thank DST, Govt. of India, for supporting this work, Nisha Johnson, Shikhar Javeri, and Hanudeep Repaka for data annotation, and **Suhas BN** for presenting.

Effect of Label Availability



- Sharp gain 0%→25%; **most gain achieved by 25%**, plateaus thereafter.
- **50% labels** ≈ full supervision.
- TVP cross-threshold consistency improves pseudo-label quality, driving strong unsupervised gains.

Results: TVP-UNet vs. Baselines

Mean (SD) of metrics (%) over 5-fold CV — *All* group

Method	Precision	Recall	F1	AUCROC
Unsupervised (0% labels)				
TVP-UNet (0%)	**84.33** (15.2)	**79.63** (17.5)	**79.30** (1.3)	68.20 (16.0)
Sarkar et al.[2]	70.55 (1.5)	71.19 (0.2)	70.86 (0.8)	**73.30** (0.5)
Supervised (100% labels)				
TVP-UNet (100%)	**96.50** (3.2)	87.86 (1.0)	**91.98** (5.1)	91.70 (2.8)
Mihalache et al.[1]	89.01 (7.2)	**93.23** (3.6)	90.75 (2.6)	**94.79** (0.5)

Mean (SD) of F1-score (%) across subgroups

Sub.	TVP (0%)	TVP (100%)	Mihalache[1]	Sarkar[2]
SV	79.86 (11.5)	**93.08** (3.8)	88.42 (2.7)	71.25 (11.0)
ML	71.55 (9.2)	**92.14** (4.1)	86.97 (2.2)	69.48 (8.7)
ND	79.27 (6.8)	**93.60** (2.7)	91.35 (1.9)	72.90 (7.5)
HC	79.57 (9.2)	88.92 (4.2)	**90.14** (3.3)	73.65 (5.5)

SV: Severe ML: Mild ND: No dysarthria HC: Healthy controls

- 0% labels: **+13.8 pp precision** and **+8.4 pp F1** vs. Sarkar et al.[2] (AUCROC −5.1 pp).
- 100% labels: **+7.5 pp precision** and +1.2 pp F1 vs. Mihalache et al.[1]
- Strong, stable F1 across SV, ML, ND subgroups.
- HC slightly lower due to class imbalance from fewer HC samples.

Conclusion

- TVP regularisation stabilises weak-estimator decisions across thresholds, enabling reliable pseudo-labelling.
- **50% labels** achieves near-supervised performance — cutting annotation cost in clinical settings.
- Consistent gains across all severity subgroups demonstrate robustness to pathological speech variability.
- **Future work:**
  - Extend to more dysarthria subtypes & multilingual datasets
  - Investigate adaptive thresholding under varying noise
  - Integrate with foundation speech models
  - Optimise for real-time and edge deployment

adityapandey.080505@gmail.com