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

ABSTRACT

We address voice activity detection (VAD) for dysarthric speech owing to neurological disorders such as Amyotrophic Lateral Sclerosis (ALS) and Parkinson's disease (PD). Atypical prosody, reduced articulatory precision, and variable intensity observed in dysarthria make standard VAD unreliable. We propose a compact U-Net autoencoder that reconstructs 100 ms waveform frames while jointly learning frame-level speech/non-speech decisions via a novel Threshold Variance Penalty (TVP). TVP stabilises a simple statistics-based weak estimator across multiple decision thresholds. TVP is combined with a reconstruction loss so that the model can be trained in supervised, semi-supervised, or unsupervised regimes. Evaluations are performed under the 0%, 25%, 50%, 75% and 100% label-availability regime with 5-fold cross-validation. The proposed method under 50% label-availability reaches a mean test F1 score of 92.46% with a mean precision of 95.59% and a mean recall of 89.57%. It performs at par with or outperforms supervised and unsupervised baselines w.r.t. the key metrics.

Index Terms— Voice activity detection; Dysarthria; Threshold Variance Penalty; Semi-supervised learning; Pseudo-labelling.

1. INTRODUCTION

Voice Activity Detection (VAD) is essential for enabling applications like Automatic Speech Recognition (ASR), speaker diarization, speech enhancement, and speech coding. Acoustic abnormalities caused by dysarthria make traditional VAD methods ineffective [1]. Dysarthria affects the muscles used for speech production [2]. This leads to reduced articulation precision, irregular prosody, variable speech rate, prolonged phonemes [3], low and highly variable intensity [4], and increased aperiodicity [5]. These changes produce concrete VAD failures, such as low-energy vowels may be dropped (long silent gaps within words), irregular/noisy phonation can be mistaken for speech (insertions), and slurring or elongations fragment word boundaries (over-/under-segmentation). Such errors degrade downstream ASR and related systems; we therefore propose a specialised VAD trained and modelled on dysarthric speech. To the best of our knowledge, no effort has yet been reported in the literature on VAD for dysarthric speech. We particularly focus on dysarthria caused by Amyotrophic Lateral Sclerosis (ALS) and Parkinson's disease (PD) in this work.

Recent supervised VAD research has increasingly adopted deep neural networks to improve detection accuracy and robustness. Transformer-based and lightweight variants (Zhao et al. [6]) exploit depthwise convolutions and audio fingerprinting for efficient modeling, while self-supervised backbones have been fine-tuned for multitask detection (Kunešová et al. [7]). Other approaches include semantic VAD that leverages punctuation prediction and transcripts (Shi et al. [8]), CNNs operating on MFCCs and prosodic cues (Mihalache et al. [9]), temporal-convolutional backbones for joint speech/overlap detection (Lebourdais et al. [10]), and SincNet encoders with spiking RNNs and attention (Liu et al. [11]). While these supervised methods exhibit strong performance and noise resilience, they typically demand substantial annotated data and complex training pipelines—an important limitation given that dysarthric speech annotation is both time-consuming and error-prone.

Despite the limited attention to unsupervised VAD, some notable methods have emerged. Bäckström et al. [12] used the Teager–Kaiser energy operator with window-overlap optimization; Tan et al.'s rVAD [13] is a two-pass, segment-level denoising detector driven by energy and SNR thresholds; Sarkar et al. [14] employed Zero-Frequency Filtering to model the vocal source and tract jointly; and Niu et al. [15] proposed quality-aware masking to purify noisy pseudo-labels during unsupervised target-speaker adaptation.

Existing VAD losses (e.g., binary cross-entropy, focal) operate at a single decision threshold and do not constrain the sensitivity of the threshold. This is problematic for dysarthric speech, where low-amplitude speech, irregular phonation, and breath noise cause speech and silence to overlap, leading

to unstable decisions. We address this using the Threshold Variance Penalty (TVP), which enforces consistency across multiple thresholds and stabilizes speech confidence under small waveform variations. Integrated with a compact U-Net, TVP enables reliable pseudo-labeling and unified supervised, semi-supervised, and unsupervised training, reducing reliance on scarce dysarthric annotations.

2. DATASET

Speech data used in this work were collected at the National Institute of Mental Health and Neurosciences (NIMHANS), Bengaluru, India. The data collection protocol was approved by the NIMHANS ethics committee, and all participants gave their written informed consent. The dataset comprises 230 ALS (161 males, 69 females), 142 PD (103 males, 39 females), and 137 HC (88 males, 49 females) subjects. The mean ages for the three subject groups are 54.53, 57.11, and 43.27 years, respectively, with the standard deviations (SD) being 11.56, 9.23, and 8.69 years, respectively. The subjects spoke thirteen native languages, including Assamese, Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Tamil, Telugu, Tulu, and Urdu. The subjects recorded spontaneous speech tasks in their respective native languages where they spoke about *a festival they celebrate* and *a place they recently visited* for $\sim$1 min each. Further details about the data collection protocol and the recording setup are available in [16]. Speech and silence segments in all recordings were manually annotated using *Audacity* [17]. Table 1 reports the count and duration details of the audio files, along with the duration statistics of the contiguous speech and silence segments present in the audio files, for each of ALS, PD, and HC populations. Three speech–language pathologists (SLPs) independently rated the dysarthria severity of each ALS and PD subject by listening to the pre-recorded spontaneous speech. Ratings were done using a 5-point scale [0, 4]. For ALS, score 0 denoted loss of useful speech and score 4 denoted normal speech, whereas the reverse was followed for PD. The mode of the three SLPs' ratings was taken as the final severity score. Table 2 shows the severity-wise count of patients and durations of audio files for ALS and PD groups. We partitioned the corpus into four groups: Severe (SV) (ALS severity score 0–1 and PD severity score 3–4), No dysarthria (ND) (ALS severity score 4 and PD severity score 0), Mild (ML) (remaining ALS and PD samples), and HC.

Table 1. Count and duration statistics of the recordings for different subject groups

Group	#Audio files	Total audio duration (min)	Mean (SD) of speech segment length (sec)	Mean (SD) of silence segment length (sec)
ALS	450	477.45	1.64 (1.10)	0.83 (1.08)
PD	278	290.65	1.73 (1.06)	1.04 (1.58)
HC	273	273.86	2.47 (2.71)	0.62 (1.02)

Table 2. Severity-wise subject count with total audio duration (in min) in brackets for ALS and PD groups

Group	Severity score				
	0	**1**	**2**	**3**	**4**
ALS	22 (40.25)	34 (70.86)	33 (72.23)	71 (145.40)	70 (148.72)
PD	76 (152.00)	50 (105.11)	14 (28.87)	1 (1.75)	1 (1.85)

3. METHODOLOGY

We propose the Temporal Variational Prior (TVP), a loss function that penalizes fluctuations in ensemble confidence across thresholds and, when combined with L_1/L_2 reconstruction losses in a U-Net–style generator (Fig. 1), yields consistent VAD decisions effective in supervised, semi-supervised, and unsupervised settings.

3.1. Preliminary

Audio is linearly normalised to $[-1, +1]$ and resampled to a common rate. Each waveform is split into non-overlapping 100-ms frames; per-waveform mean and standard deviation are computed for use in the loss terms (see Section 3.3). Ground-truth annotations are synchronised to the resampled signal and sliced into corresponding 100-ms binary labels ($0 =$ silence, $1 =$ speech).

3.2. Model Architecture

We use a 1D U-Net–style autoencoder on raw audio inspired by classical 2D U-Net [18], with initial dropout (p=0.3), three max-pool downsampling stages, and a bottleneck with layer normalisation [19], additive noise sampled from standard gaussian distribution, and a learned multiplicative gate (tanh/sigmoid with SiLU) shown in Fig. 2. The encoder is asymmetric (three down blocks) to the decoder having four convolutional stages $4f \to 2f$, $2f \to f$, $f \to f$, and a final Conv1D $\to C_{in}$, where f is the hidden-channel dimension initialized to value of 6 and C_{in} is the input-channel dimension. Nearest-neighbour upsampling, additive skip connections,

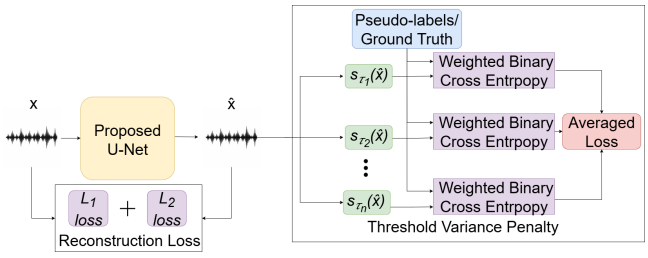

Fig. 1. Proposed training mechanism; here, x: input signal, $\hat{x}$: reconstructed signal, s_{τ_i}: weak estimator, τ_i: i^{th} threshold

and GELU activations are used. Training is end-to-end with reconstruction loss plus TVP to produce an augmented waveform that reduces VAD decision variability (Section 3.3).

3.3. Learning Objective

The learning objective includes two loss functions, Reconstruction Loss ($\mathcal{L}_{rec}$), and TVP ($\mathcal{L}_{tvp}$). The total loss ($\mathcal{L}_{total}$) is given by a weighted sum of these two losses as,

$$\mathcal{L}_{total}(x, \hat{x}) = \alpha \mathcal{L}_{rec}(x, \hat{x}) + (1 - \alpha)\mathcal{L}_{tvp}(\hat{x}). \quad (1)$$

Here, x is the input audio signal, $\hat{x}$ is the reconstructed signal, and α is a scalar hyperparameter satisfying $\alpha \in [0, 1]$ that controls the trade-off between reconstruction and TVP (i.e., $\alpha = 1$ yields pure reconstruction training, while $\alpha = 0$ relies solely on the TVP). We initialize $\alpha = 1.0$ and total epochs (E) as 100, set decay rate $\rho = 0.9$ and $\alpha_{\min} = 0.4$. α is constant within each epoch. After the training midpoint (epochs $> \lceil E/2 \rceil$), it is multiplied by ρ at each epoch end and clipped to be no less than $\alpha_{\min}$.

3.3.1. Reconstruction Loss

For Reconstruction loss, a linear combination of both L_1 and L_2 losses is applied between the input frames and their reconstructions. The L_1 term promotes sparsity and sharper reconstructions, while the L_2 term ensures stability by penalising large deviations.

3.3.2. Threshold Variance Penalty

We compute the TVP from a simple statistical weak estimator that flags likely speech samples by hard decision based on the mean and standard deviation. For an speech sample $\hat{x}$ with mean $\mu_{\hat{x}}$ and standard deviation $\sigma_{\hat{x}}$, the weak estimator $\tilde{y}(\hat{x})$ is defined as:

$$\tilde{y}(\hat{x}, \tau) = \begin{cases} 0, & \text{if } \dfrac{|\hat{x} - \mu_{\hat{x}}|}{\sigma_{\hat{x}}} \leq \tau, \\ 1, & \text{otherwise.} \end{cases} \quad (2)$$

where τ is a scalar hyperparameter threshold in range $[0, \infty)$. To allow end-to-end gradient propagation, we replace Eq. 2 by a smooth approximation:

$$s_\tau(\hat{x}) \;=\; \sigma\Big(\frac{|\hat{x} - \mu_{\hat{x}}|}{\sigma_{\hat{x}}} \;-\; \tau\Big), \quad (3)$$

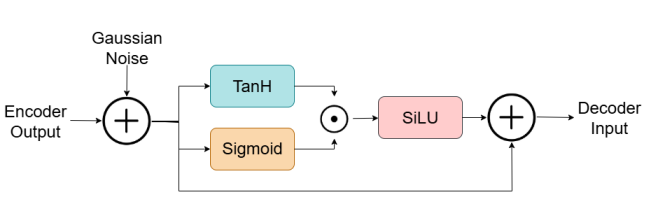

Fig. 2. Latent space processing in the proposed U-Net

where $\sigma(\cdot)$ is the sigmoid activation function. The scalar $(s_\tau(\hat{x}) \in (0, 1))$ is a differentiable soft score that approximates the binary decision at the threshold τ. Given a set of discrete threshold $\mathcal{T} = \{\tau_1, \ldots, \tau_K\}$, we define the mean threshold $\bar{\tau}$ as arithmetic mean of all thresholds in set $\mathcal{T}$. For unlabeled frames, we form a pseudo-label by applying the equation (Eq. 2) at the mean threshold $\bar{\tau}$. The $\tilde{y}(\hat{x})$ is treated as a fixed target (stop gradient) pseudo-label in loss computation. To compensate for class imbalance, we use Weighted Binary Cross Entropy (L_{WBCE}) as defined in Das et al. [20]. The weight β is updated per epoch via

$$\beta_t \leftarrow \theta \beta_{t-1} + (1 - \theta)\frac{N_{\text{neg}}}{N_{\text{pos}} + \varepsilon}, \quad (4)$$

with $\theta = 0.8$, $\varepsilon = 10^{-6}$ and initial $\beta_0 = 1.0$. Aggregated counts of total instances of speech class (N_{pos}) and silence class (N_{neg}) are computed across the entire epoch using ground-truth labels where available and the hard pseudo-labels for unlabeled frames. A sample is called label-available if a ground-truth label $y \in \{0, 1\}$ exists for the respective sample. For such labelled available samples, we compute the supervised TVP by averaging weighted BCE across thresholds:

$$\mathcal{L}_{\text{tvp}}^{\text{sup}}(\hat{x}) \;=\; \frac{1}{K}\sum_{k=1}^{K} \mathrm{L}_{WBCE}\big(s_{\tau_k}(\hat{x}),\, y\big). \quad (5)$$

For non-label-available frames, we enforce consistency between each soft threshold response and the hard pseudo-label:

$$\mathcal{L}_{\text{tvp}}^{\text{unsup}}(\hat{x}) \;=\; \frac{1}{K}\sum_{k=1}^{K} \mathrm{L}_{WBCE}\big(s_{\tau_k}(\hat{x}),\, \tilde{y}(\hat{x})\big), \quad (6)$$

Combining both cases, with $\mathbb{I}_{\text{lab}}(x)$ indicating label-availability of a frame x, the TVP per sample is

$$\mathcal{L}_{\text{tvp}}(\hat{x}) \;=\; \mathbb{I}_{\text{lab}}(x)\, \mathcal{L}_{\text{tvp}}^{\text{sup}}(\hat{x}) \;+\; \big(1 - \mathbb{I}_{\text{lab}}(x)\big)\, \mathcal{L}_{\text{tvp}}^{\text{unsup}}(\hat{x}), \quad (7)$$

and the final TVP used in the training is the mean of $\mathcal{L}_{\text{tvp}}(\hat{x})$ across the mini-batch.

4. EXPERIMENTAL SETUP

All experiments of our proposed method used five label-availabilities $0\%, 25\%, 50\%, 75\%, 100\%$. All experiments were carried out with Adam optimiser (learning rate = 1×10^{-3}) for a fixed number of $E = 100$ epochs and batch size $B = 32$. All convolutional layers use a kernel size of 3 (with the final reconstruction layer using a kernel size of 1), a stride 1, and no bias term. The max-pooling layer is applied with a stride 2, while the upsampling layer uses a scale factor of 2. The similarity term uses a user-defined threshold set $\mathcal{T} = [0.05,\ 0.1,\ 0.15]$. The mean threshold $\bar{\tau} = 0.1$ is used to produce hard pseudo-labels for unlabeled frames using Eq. 2. Evaluation metrics reported are precision, recall, F1-score,

and AUCROC. For baselines, we use [9] and [14]. Results are averaged over a five-fold cross-validation, where each fold ensures that speakers are disjoint in the training and test sets. Both baselines are trained from scratch on our dataset using the same protocol. For each test fold, metrics are computed per audio file, then aggregated into four groups (SV, ML, ND, HC), with the mean and standard deviation reported for each. Means and standard deviations computed across all test folds are reported under the *All* group.

5. RESULTS

Across five-fold cross-validation (Table 3), TVP-UNet shows clear advantages in both unsupervised and fully supervised regimes. In the unsupervised setting (0% labels), the model substantially improves precision (+13.8 percentage points(pp)) and yields notable gains in recall/F1 (approximately +8.4 pp) relative to the unsupervised baseline (Sarkar et al. [14]), at the cost of a modest decrease in AUCROC (−5.1 pp). Under full supervision (100% labels), TVP-UNet maintains superior precision (+7.49 pp) and a small F1 advantage (+1.23 pp) versus the supervised baseline (Mihalache et al. [9]), indicating better performance both when labels are scarce and when labels are abundant. Performance is consistent across dysarthric subtypes (SV, ML, ND). TVP-UNet achieves high and stable F1-scores for these groups in the supervised regime and preserves substantial gains in the unsupervised regime, demonstrating robustness to different manifestations of dysarthria. By contrast, the HC group exhibits lower performance, particularly in AUCROC and recall. This degradation is most likely attributable to the small number of HC samples and resulting class imbalance, which reduces statistical reliability across folds.

Figure 3 shows mean and standard deviation across folds for F1, Recall, Precision, and AUCROC at different label fractions (0–100%). Performance improves markedly from 0% to 50% labels and then largely plateaus, with much of the gain and variance reduction already achieved at 25%. Precision increases nearly monotonically, while recall shows moderate variability at intermediate levels. About 50% labeled data attains performance close to full supervision while remaining stable, which is important given the labor-intensive nature of dysarthria annotation. The strong unsupervised performance stems from reduced threshold sensitivity: enforcing consistency stabilizes decisions in overlapping speech–silence regions and improves pseudo-label quality.

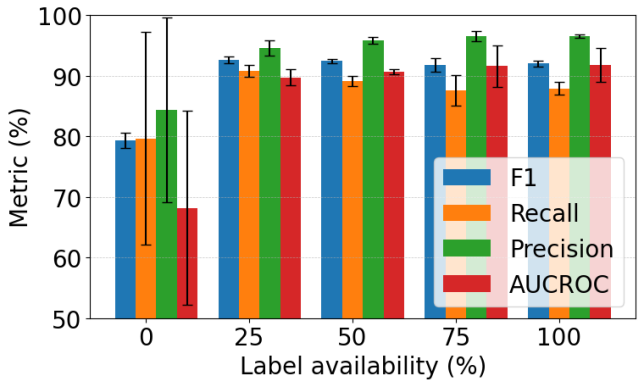

Fig. 3. Mean and Standard deviation (as error bar) of the proposed method for different % of label availability

Table 3. Mean (SD in bracket) of evaluation metrics in % obtained over 5-fold cross-validation for different subsets of the test set using the proposed and baseline methods

Method	Group	Precision	Recall	F1-score	AUCROC
TVP-UNET (0%)	SV	81.40(20.2)	84.80(13.5)	79.86(11.5)	71.68(17.1)
	ML	81.76(19.2)	80.66(15.8)	71.55(9.2)	69.80(15.6)
	ND	83.74(16.5)	80.90(16.0)	79.27(6.8)	68.51(15.3)
	HC	88.64(12.2)	76.78(19.3)	79.57(9.2)	65.50(13.2)
	All	84.33(15.2)	79.63(17.5)	79.3(1.3)	68.20(16.0)
TVP-UNET (100%)	SV	96.48(5.0)	90.18(5.1)	93.08(3.8)	92.89(3.6)
	ML	95.65(5.4)	89.20(5.6)	92.14(4.1)	91.20(5.7)
	ND	96.90(3.5)	90.74(4.6)	93.60(2.7)	92.41(3.3)
	HC	96.41(6.2)	88.93(4.2)	88.92(4.2)	86.27(7.0)
	All	96.50(3.2)	87.86(1.0)	91.98(5.1)	91.70(2.8)
Mihalache et al. [9]	SV	93.70(5.6)	91.93(4.6)	92.28(2.7)	97.36(0.7)
	ML	93.05(4.5)	90.81(6.3)	91.49(2.2)	96.57(0.7)
	ND	94.29(4.6)	91.38(6.1)	92.43(1.9)	97.09(0.3)
	HC	95.45(3.2)	83.45(7.9)	88.42(3.3)	92.54(0.5)
	All	89.01(7.2)	93.23(3.6)	90.75(2.6)	94.79(0.5)
Sarkar et al. [14]	SV	60.47(17.2)	66.75(5.8)	61.52(11.0)	52.14(3.3)
	ML	65.30(14.3)	71.72(2.9)	67.38(8.7)	54.29(2.5)
	ND	70.11(12.9)	73.11(1.9)	70.85(7.5)	65.29(2.2)
	HC	81.19(11.6)	70.95(2.6)	75.24(5.5)	74.89(3.2)
	All	70.55(1.5)	71.19(0.2)	70.86(0.8)	73.30(0.5)

6. CONCLUSION

We introduced TVP, a simple and effective regularizer that stabilizes weak-estimator decisions across thresholds and, when combined with reconstruction losses in a compact U-Net, yields consistent VAD for dysarthric speech. The method achieves a favorable precision–recall tradeoff and competitive AUCROC while reducing the reliance on labeled data, making it well-suited for clinical and low-resource scenarios. Future work includes extending evaluations to additional dysarthria subtypes and languages, and integrating the model into a real-time front end for downstream ASR and clinical monitoring.

Acknowledgements - We thank the Department of Science and Technology (DST), Govt. of India, for supporting this work. We also thank Nisha Johnson, Shikhar Javeri, and Hanudeep Repaka for their help in data annotation.

7. REFERENCES

[1] Saeid Alavi Naeini, Leif Simmatis, Deniz Jafari, Yana Yunusova, and Babak Taati, "Improving dysarthric speech segmentation with emulated and synthetic augmentation," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 12, pp. 382–389, 2024.

[2] American Speech-Language-Hearing Association, "Dysarthria in adults," `https://www.asha.org/practice-portal/clinical-topics/dysarthria-in-adults/`.

[3] Austin Thompson and Yunjung Kim, "Acoustic and kinematic predictors of intelligibility and articulatory precision in parkinson's disease," *Journal of Speech, Language, and Hearing Research*, vol. 67, no. 10, pp. 3595–3611, 2024.

[4] Caitlin Cloud, Kaily Georgen-Schwartz, and Allison Hilger, "The contributions of pitch, loudness, and rate control to speech naturalness in cerebellar ataxia," *American Journal of Speech-Language Pathology*, vol. 33, no. 5, pp. 2536–2555, 2024.

[5] Noé Xiu, Wenmei Li, Lu Liu, Zhaoqi Liu, Zhuo Cai, Lanlan Li, Béatrice Vaxelaire, Rudolph Sock, Zhenhua Ling, Juluo Chen, and Youmeng Wang, "A study on voice measures in patients with parkinson's disease," *Journal of Voice*, 2024 (In press).

[6] Yifei Zhao and Benoit Champagne, "An efficient transformer-based model for voice activity detection," in *IEEE* 32^{nd} *International Workshop on Machine Learning for Signal Processing (MLSP)*, 2022, pp. 1–6.

[7] Marie Kunešová and Zbyněk Zajíc, "Multitask detection of speaker changes, overlapping speech and voice activity using wav2vec 2.0," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[8] Mohan Shi, Yuchun Shu, Lingyun Zuo, Qian Chen, Shiliang Zhang, Jie Zhang, and Li-Rong Dai, "Semantic vad: Low-latency voice activity detection for speech interaction," in *Proc. Interspeech*, 08 2023, pp. 5047–5051.

[9] Serban Mihalache and Dragos Burileanu, "Using voice activity detection and deep neural networks with hybrid speech feature extraction for deceptive speech detection," *Sensors*, vol. 22, no. 3, 2022.

[10] Martin Lebourdais, Théo Mariotte, Marie Tahon, Anthony Larcher, Antoine Laurent, Silvio Montresor, Sylvain Meignier, and Jean-Hugh Thomas, "Joint speech and overlap detection: a benchmark over multiple audio setup and speech domains," *arXiv preprint arXiv:2307.13012*, 2023.

[11] Qu Yang, Qianhui Liu, Nan Li, Meng Ge, Zeyang Song, and Haizhou Li, "Svad: A robust, low-power, and light-weight voice activity detection with spiking neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 221–225.

[12] Tom Bäckström, "Overlap-add windows with maximum energy concentration for speech and audio processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 491–495.

[13] Zheng-Hua Tan, Achintya kr. Sarkar, and Najim Dehak, "rvad: An unsupervised segment-based robust voice activity detection method," *Computer Speech & Language*, vol. 59, pp. 1–21, 2020.

[14] Eklavya Sarkar, Ravishankar Prasad, and Mathew Magimai-Doss, "Unsupervised voice activity detection by modeling source and system information using zero frequency filtering," in *Proc. Interspeech*, 09 2022, pp. 4626–4630.

[15] Shutong Niu, Jun Du, Maokui He, Chin-Hui Lee, Baoxiang Li, and Jiakui Li, "Unsupervised adaptation with quality-aware masking to improve target-speaker voice activity detection for speaker diarization," in *Proc. Interspeech*, 08 2023, pp. 3482–3486.

[16] Jhansi Mallela, Yamini Belur, Nalini Atchayaram, Ravi Yadav, Pradeep Reddy, Dipanjan Gope, and Prasanta Kumar Ghosh, "Raw speech waveform based classification of patients with ALS, Parkinson's disease and healthy controls using CNN-BLSTM," in *Proc. Interspeech*, 2020, pp. 4586–4590.

[17] Audacity Team, "Audacity: Free, open source, cross-platform software for recording and editing sounds [computer program], version 3.1.0," retrieved from https://www.audacityteam.org, 2020.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI*, Cham, 2015, pp. 234–241, Springer International Publishing.

[19] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[20] Nilanjana Das and Shahin Ara Begum, "An empirical study of loss functions for aspect category detection in imbalanced data scenario," in 10^{th} *International Conference on Signal Processing and Communication (ICSC)*, 2025, pp. 247–252.