ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446945

SVAD: A ROBUST, LOW-POWER, AND LIGHT-WEIGHT VOICE ACTIVITY DETECTION WITH SPIKING NEURAL NETWORKS

Qu Yang[1*], *Qianhui Liu*[1*✉], *Nan Li*[2], *Meng Ge*[1], *Zeyang Song*[1], *Haizhou Li*[3,1]

[1] National University of Singapore, Singapore [2] Tianjin University, Tianjin, China
[3] The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China

ABSTRACT

Speech applications are expected to be low-power and robust under noisy conditions. An effective Voice Activity Detection (VAD) front-end lowers the computational need. Spiking Neural Networks (SNNs) are known to be biologically plausible and power-efficient. However, SNN-based VADs have yet to achieve noise robustness and often require large models for high performance. This paper introduces a novel SNN-based VAD model, referred to as sVAD, which features an auditory encoder with an SNN-based attention mechanism. Particularly, it provides effective auditory feature representation through SincNet and 1D convolution, and improves noise robustness with attention mechanisms. The classifier utilizes Spiking Recurrent Neural Networks (sRNN) to exploit temporal speech information. Experimental results demonstrate that our sVAD achieves remarkable noise robustness and meanwhile maintains low power consumption and a small footprint, making it a promising solution for real-world VAD applications.

Index Terms— Voice Activity Detection (VAD), Spiking Neural Network (SNN), Auditory attention, Noise robustness

1. INTRODUCTION

Voice Activity Detection (VAD) plays a pivotal role as a front-end in various speech applications such as automatic speech recognition, keyword spotting [1, 2, 3]. It detects whether the speech is present in an audio signal to activate subsequent applications only when speech is detected. This paradigm helps to save computational resources and enhance the overall efficiency of the system [4].

VAD serves as an always-on model that is required to operate efficiently with low power consumption. Simultaneously, as a front-end, it needs to be light-weight, ensuring minimal memory utilization while maintaining its performance. Beyond these prerequisites, noise robustness is also critical so that the VAD can operate stably under different noise levels. Hence, there is a demand for a robust, low-power, and light-weight VAD model.

Spiking Neural Networks (SNNs) mimic the information processing mechanism in the human brain [5, 6, 7, 8]. A spiking neuron is only active when it receives or emits a spike, facilitating power-efficient processing. Incoming spikes increment the neuron's membrane potential through accumulate (AC) operations, a process more power-efficient than the multiply-accumulate (MAC) operations in traditional artificial neural networks (ANNs) [9, 10, 11, 12, 13]. This intrinsic power-saving characteristic renders SNNs as optimal candidates for VAD applications.

However, SNN-based VAD suffers a significant performance loss under noisy conditions. It was reported that system performance in low speech-to-noise (SNR) ratio can be 25% lower than that in high SNR conditions [14]. Additionally, achieving optimal performance often necessitates large SNN models. [3] demonstrated the best SNN-based VAD results using a model containing over 1M parameters. Such a large model can be impractical in many real-world applications. [14] employed a compact 2.6K-parameter model, which regrettably shown poor performance.

To address these limitations, we propose a novel SNN-based VAD model, referred to as sVAD, that is expected to be noise-robust, low-power, and light-weight. We present an auditory encoder integrated with an SNN-based attention mechanism. This encoder employs both SincNet [15] and 1D convolution to extract flexible and interpretable auditory features in a data-driven manner, thereby resulting in effective feature representation and enhancing the overall performance. The incorporation of the SNN-based attention mechanism improves the saliency of the extracted features, thus increasing the robustness of VAD model. The classifier employs the Spiking Recurrent Neural Networks (sRNN) that can exploit the temporal information contained in the speech. Experimental results demonstrate our proposed sVAD achieves remarkable noise robustness and meanwhile maintains low power consumption and a small footprint.

The rest of the paper is organized as follows. Section 2 presents the SNN-based VAD incorporating with auditory attention. Section 3 reports our experiments and results. Finally, we conclude the work in Section 4.

*Equal Contribution. ✉ Corresponding author. qhliu@nus.edu.sg
This work is supported by IAF, A*STAR, SOITEC, NXP and National University of Singapore under FD-fAbrICS: Joint Lab for FD-SOI Always-on Intelligent & Connected Systems (Award I2001E0053).

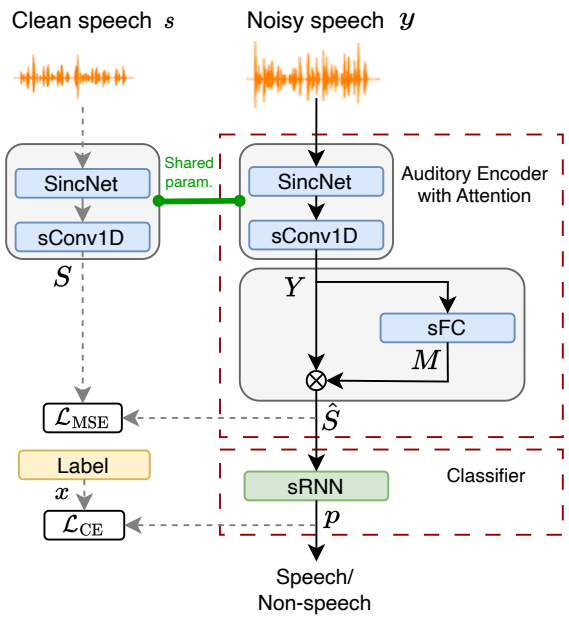

Fig. 1. The proposed SNN-based VAD model consists of an auditory encoder with attention for feature extraction and a classifier for frame-level classification.

2. SNN-BASED VAD WITH AUDITORY ATTENTION

As shown in Fig. 1, the proposed SNN-based VAD model consists of two key components: 1) an auditory encoder with attention, which converts raw audio input into spike-featured frames; and 2) a classifier dedicated to the frame-by-frame voice activity detection task. The VAD model benefits from the energy efficiency of event-driven computation of SNN.

2.1. Spiking Neuron Model

We employ the widely recognized Leaky Integrate-and-Fire (LIF) model [16] for spiking neurons. LIF neurons integrate synaptic inputs until their membrane potential exceeds a firing threshold, resulting in output spikes for transmission to subsequent neurons. The LIF neuron's dynamics are captured by the following discrete-time expressions:

$$U_i^l[t] = \alpha U_i^l[t-1] + I_i^l[t] - \vartheta O_i^l[t-1], \tag{1}$$

$$I_i^l[t] = \sum_j w_{ij}^{l-1} O_j^{l-1}[t-1] + b_i^l, \tag{2}$$

where $U_i^l[t]$ represents the membrane potential of neuron i at layer l, and $I_i^l[t]$ denotes the input current at time t. α signifies the membrane potential decay constant, while ϑ represents the neuronal firing threshold. w_{ij}^{l-1} denotes the connection weight from neuron j in the preceding layer $l-1$, and b_i^l is the constant current injected into neuron i. The occurrence of an output spike, denoted as $O_i^l[t-1]$, is determined using the spiking activation function:

$$O_i^l[t] = \Theta\left(U_i^l[t] - \vartheta\right). \tag{3}$$

Here, $\Theta(\cdot)$ is the Heaviside step function.

2.2. Auditory Encoder with Attention

2.2.1. *Auditory encoder*

Our study leverages the well-established auditory encoder, SincNet [15], known for its exceptional feature extraction capabilities. In contrast to conventional Convolutional Neural Networks (CNNs) processing raw speech data, SincNet introduces constraints on filter shapes in its first layer, making it highly suitable for our VAD model. It uses parametrized band-pass filters with only two parameters, ensuring interpretability and human intuition in the feature extraction process. By integrating SincNet into our auditory encoder, we enhance feature extraction, improving the performance of our SNN-based VAD model across various noise conditions while maintaining a light-weight configuration.

In general, 1D convolutional layers introduce trainable parameters enhancing encoder adaptability to data characteristics and improving performance through training. Inspired by ConvTasNet [17], we extend this data-driven ability by adding an SNN-based 1D convolutional layer (sConv1D) to SincNet-processed features. This empowers the encoder to optimize auditory feature extraction during training. Additionally, sConv1D outputs spikes, facilitating the transformation of raw audio into spike representation for subsequent SNN-based processing.

2.2.2. *Attention mechanism*

To enhance our VAD model robustness, especially in low SNR conditions, we introduce an attention mechanism inspired by the human auditory system's masking effect. In the encoder module (shown in Fig. 1), we implement this attention mechanism as follows: Initially, we subject the extracted features Y to three layers of SNN-based fully connected (sFC) layers to derive an attention mask M. Subsequently, we leverage this attention mask to modulate the extracted features, generating attended features denoted as $\hat{S}$:

$$\hat{S} = Y \odot M \tag{4}$$

where $\odot$ is the element-wise multiplication. The trainable three-layer sFC block optimizes the attention mask for different acoustic conditions.

2.3. Model Learning

2.3.1. *Spiking recurrent neural network for classification*

The classifier utilizes the encoded features as its input and is trained to perform frame-by-frame decision-making dur-

ing runtime. To enhance the sequential modeling capacity, we establish the classifier using SNN with recurrent connections, referred to as sRNN as shown in Fig. 1. Unlike the feedforward processing update formula (Eq. (2)), the update equation for recurrent neurons includes an additional term for recurrent connections:

$$I_i^l[t] = \sum_j w_{ij}^{l-1} O_j^{l-1}[t-1] + \sum_i w_{ii}^l O_i^l[t-1] + b_i^l, \quad (5)$$

where w_{ii}^l denotes the recurrent connection weight of neuron i in layer l.

2.3.2. *Loss function*

The proposed VAD model's loss function comprises two components: a classification loss and an attention mask loss. For the classification loss, we employ cross-entropy (CE):

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{2} x_c \log(p_c). \quad (6)$$

Here, $[x_1, x_2]$ denotes an one-hot encoding ($[0, 1]$ for speech and $[1, 0]$ for non-speech), and p_c is the Softmax probability that the input belongs to class c. To calculate the attention mask loss, we utilize the mean squared error (MSE):

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{k=1}^{N} (\hat{S}_k - S_k)^2, \quad (7)$$

where S is the clean speech features, and $\hat{S}$ is the modulated features described in Eq. (4). N denotes the total number of elements in the features and k represents the index of elements. Consequently, the overall loss of our proposed sVAD model is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{MSE}}, \quad (8)$$

where λ is the hyperparameter for balancing both losses.

2.3.3. *SNN training algorithm*

The stateful nature of spiking neurons, similar to vanilla RNNs, enables SNNs training by unfolding the network across all time steps and employing the Backpropagation Through Time (BPTT) algorithm [18]. However, the non-differentiable spiking activation function, as outlined in Eq. 3, poses a challenge for direct BPTT application. This study addresses it by adopting the surrogate gradient approach [19], introducing a boxcar function as an effective surrogate gradient:

$$\Theta'(U_i^l[t] - \vartheta) \approx \theta'(U_i^l[t] - \vartheta) = \frac{1}{a} sign\left(|U_i^l[t] - \vartheta| < \frac{a}{2}\right), \quad (9)$$

where the hyperparameter a controls the permissible range of membrane potentials that allow gradients to pass through.

3. EXPERIMENTS AND RESULTS

3.1. Dataset and setup

We evaluate the proposed sVAD model on the QUT-NOISE-TIMIT dataset [20], consisting of 600 hours of noisy speech. This dataset combines clean speech recordings from the TIMIT dataset with real-world noise scenarios, such as cafe, car, home, street, and reverberant environments. The dataset is categorized into three noise levels: low (SNR = +15dB, +10dB), medium (SNR = +5dB, 0dB), and high (SNR = -5dB, -10dB), with further stratification based on noise environment (Group A and Group B). Training and testing alternate between these groups, covering all noise levels. For the evaluation metrics, we report results using the frame-level Half-Total Error Rate (HTER), averaging the Miss Rate (MR) and False Alarm Rate (FAR).

Feature extraction by the auditory encoder produces 20-dimensional features with a $30ms$ frame size and 50% overlap, enabling $15ms$ frame-by-frame classification and achieving a low latency of $15ms$ for our models. The classifier comprises 32 hidden recurrent spiking neurons and 2 linear readout neurons for the "speech" and "non-speech" classes. Spiking neuron hyperparameters are set to $\alpha = 0.5$, $\vartheta = 0.3$, and $a = 4$.

All experiments use PyTorch with GPU acceleration, employing the Adam optimizer for 100 epochs, a batch size of 128, and an initial learning rate of 0.001. The learning rate reduces it by a factor of ten every 40 epochs.

3.2. Results and analysis

3.2.1. *Compare with existing models*

We evaluate our sVAD model by comparing it with ten baseline systems, including standards-based approaches (e.g., ETSI [21] and G729B [22]), statistical methods (e.g., LTSD [23] and Sohn [24]), machine learning-based solutions (e.g., GMM [20], CLC [25], and CNN [26]), and recent SNN-based VAD models (e.g., Bin e. [14], SNN h1/h1-p [27], and VAD system in HuRAI [3]).

For a comprehensive and equitable comparison, we first assess all SNN-based VAD models, considering parameters count and HTER across different noise levels (low, medium, and high). Results are reported in Table 1. Notably, the VAD in HuRAI employs significantly more parameters, making it impractical for real-world use [4]. To ensure fairness, we reconduct the VAD in HuRAI by aligning the parameter count with those used in our model. For a comparison with the Bin e. model, which uses only 2.6K parameters, we create a reduced version, named sVAD-S, with a smaller sRNN layer of just 10 recurrent spiking neurons and two sFC layers for attention mask estimation. Despite these reductions, our modified model exhibits only a marginal increase in HTER across all three noise levels. In addition, it is worth mentioning that

Table 1. A summary of SNN-based VAD models, including parameter count and HTER across noise levels (Low, Medium, and High). * denotes our reproduced results.

Model	# Param.	Low (%)	Medium (%)	High (%)
HuRAI [3]	>1M	2.7	6.7	15.0
HuRAI* [3]	4.4K	9.8	23.4	25.0
Bin e. [14]	2.6K	8.2	23.6	33.6
SNN h1 [27]	26K	4.6	12.4	25.2
SNN h1-p [27]	4.1K	4.7	12.5	25.8
sVAD	4.3K	4.0	11.9	19.1
sVAD-S	2.4K	5.8	12.6	22.3

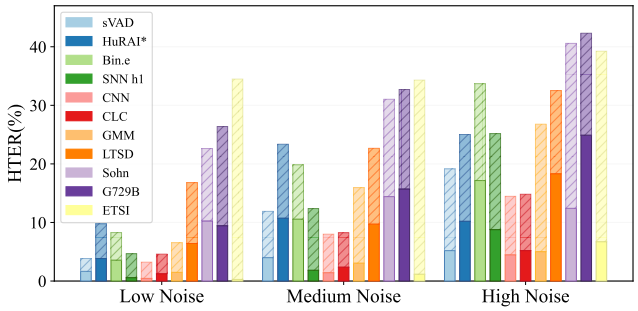

Fig. 2. Comparisons of HTER performance for different noise levels. Dark and light shadings denote the miss rate and false alarm rate contributions, respectively.

our models use the smallest frame shift (i.e., $15ms$) for the encoder among all SNN-based models (e.g., $20ms$ in [14]), resulting in the lowest latency among them.

Next, we compare our sVAD model with the aforementioned ten baseline systems, as depicted in Fig. 2. The results demonstrate that our sVAD outperforms most models, with the exception of the CNN and CLC models. This can be attributed to the fact that the CNN and CLC models use larger frame sizes and shifts for decision-making (with frame sizes and shifts more than 5 times for CNN and more than 3 times for CLC in comparison to our models), thereby introducing increased latency. Despite this discrepancy, it's crucial to emphasize that the primary focus of this study is the development of a robust, low-power, and light-weight SNN-based VAD model under noisy conditions rather than striving for the top HTER performance.

Overall, our model stands out among SNN-based VAD models, delivering superior performance, especially in high-noise conditions, while maintaining a competitive parameter count. This achievement signifies the successful development of a robust, low-power, and light-weight SNN-based VAD model under noisy conditions.

3.2.2. Ablation study

To validate the efficacy of our proposed auditory encoder, we conduct an ablation study on high-noise conditions, with results detailed in Table 2. We employ the proposed auditory encoder used in sVAD as the baseline. First, we remove the learnable sConv1D layer from the encoder, resulting in a significant 4.1% increase in HTER. Subsequently, the further elimination of the attention mechanism led to another noteworthy 4.0% increase in HTER. In summary, our ablation study demonstrates compelling evidence for the efficacy of our auditory encoder in significantly improving the robustness of our SNN-based VAD model under noisy conditions.

Table 2. Ablation study on the proposed Auditory Encoder's efficacy, focusing on high-noise conditions (SNR: -5dB and -10dB).

Encoder Model	MR	FAR	HTER
sVAD	10.4%	27.9%	19.1%
- sConv1D	19.5%	26.9%	23.2%
- sConv1D & Attention	24.6%	29.8%	27.2%

Table 3. Power consumption of our sVAD and other baselines.

Model	Power Consumption
sVAD / sVAD-S	2.0 / 0.9μW (lower bound)
Bin e. [14]	3.8 μW (lower bound)
SNN h1-p [27]	25.1 μW
Yang et al [28]	1 μW
Price et al [2]	22 μW
Meoni et al [29]	559 μW

3.2.3. Power consumption

This section estimates the power consumption of our sVAD models and compares them with that of other VADs. Our estimation is based on the Loihi neuromorphic chip [30] and follows the estimation methodology in [14], encompassing power consumption from synaptic operations and neuron updates. As depicted in Table 3, our sVADs demonstrate lower power consumption, particularly the smaller sVAD-S model. It is worth noting that we only estimate the lower bound based on Loihi, which is different from [28, 2, 29] that have run on ASIC chips.

4. CONCLUSION

We develop a novel SNN-based VAD mode consisting of an auditory encoder with attention for feature extraction and an sRNN for classification. The auditory encoder with attention can provide an effective and robust feature representation of the raw audio, which enables us to deploy a light-weight SNN while still maintaining competitive performance. Furthermore, the model's small footprint, coupled with the energy-efficient characteristics of SNN, results in low power consumption. Comparison experimental results with other VADs show that our proposed sVAD has high noise robustness, low power consumption, and a small footprint. The ablation study further validates the effectiveness and robustness of our proposed auditory encoder with attention.

5. REFERENCES

[1] S. Oh, M. Cho, Z. Shi, J. Lim, Y. Kim, S. Jeong, Y. Chen, R. Rothe, D. Blaauw, H.-S. Kim *et al.*, "An acoustic signal processing chip with 142-nw voice activity detection using mixer-based sequential frequency scanning and neural network classification," *IEEE JSSC*, vol. 54, no. 11, pp. 3005–3016, 2019.

[2] M. Price, J. Glass, and A. P. Chandrakasan, "A low-power speech recognizer and voice activity detector using deep neural networks," *IEEE JSSC*, vol. 53, no. 1, pp. 66–75, 2017.

[3] J. Wu, Q. Liu, M. Zhang, Z. Pan, H. Li, and K. C. Tan, "Hurai: A brain-inspired computational model for human-robot auditory interface," *Neurocomputing*, vol. 465, pp. 103–113, 2021.

[4] S. Yadav, P. A. D. Legaspi, M. S. O. Alink, A. B. Kokkeler, and B. Nauta, "Hardware implementations for voice activity detection: trends, challenges and outlook," *IEEE TCAS-I*, vol. 70, no. 3, pp. 1083–1096, 2022.

[5] Q. Liu, H. Ruan, D. Xing, H. Tang, and G. Pan, "Effective aer object classification using segmented probability-maximization learning in spiking neural networks," in *AAAI*, vol. 34, no. 02, 2020, pp. 1308–1315.

[6] Q. Liu, D. Xing, L. Feng, H. Tang, and G. Pan, "Event-based multimodal spiking neural network with attention mechanism," in *ICASSP*. IEEE, 2022, pp. 8922–8926.

[7] Q. Yang, Q. Liu, and H. Li, "Deep residual spiking neural network for keyword spotting in low-resource settings," *Interspeech 2022*, pp. 3023–3027, 2022.

[8] Q. Yang, J. Wu, M. Zhang, Y. Chua, X. Wang, and H. Li, "Training spiking neural networks with local tandem learning," *NeurIPS*, vol. 35, pp. 12 662–12 676, 2022.

[9] C. Farabet, R. Paz, J. Pérez-Carrasco, C. Zamarreño-Ramos, A. Linares-Barranco, Y. LeCun, E. Culurciello, T. Serrano-Gotarredona, and B. Linares-Barranco, "Comparison between frame-constrained fix-pixel-value and frame-free spiking-dynamic-pixel convnets for visual processing," *Front. Neurosci.*, vol. 6, p. 32, 2012.

[10] X. Ma, G. Fang, and X. Wang, "LLM-Pruner: On the Structural Pruning of Large Language Models," in *NeurIPS*, 2023.

[11] G. Fang, X. Ma, M. Song, M. B. Mi, and X. Wang, "DepGraph: Towards Any Structural Pruning," in *CVPR*, 2023, pp. 16 091–16 101.

[12] X. Ma, G. Fang, and X. Wang, "Deepcache: Accelerating diffusion models for free," *arXiv preprint arXiv:2312.00858*, 2023.

[13] X. Yang, D. Zhou, S. Liu, J. Ye, and X. Wang, "Deep Model Reassembly," in *NeurIPS*, vol. 35, 2022, pp. 25 739–25 753.

[14] G. Dellaferrera, F. Martinelli, and M. Cernak, "A bin encoding training of a spiking neural network based voice activity detection," in *ICASSP*. IEEE, 2020, pp. 3207–3211.

[15] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *SLT workshop*. IEEE, 2018, pp. 1021–1028.

[16] W. Gerstner and W. M. Kistler, *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.

[17] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM TASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[18] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[19] E. O. Neftci, H. Mostafa, and F. Zenke, "Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks," *IEEE Signal Processing Mag.*, vol. 36, no. 6, pp. 51–63, 2019.

[20] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The qut-noise-timit corpus for evaluation of voice activity detection algorithms," in *Interspeech*, 2010, pp. 3110–3113.

[21] E. S. Doc, "Speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI ES*, vol. 202, no. 050, p. v1, 2002.

[22] A. Benyassine, E. Shlomot, H. Su, D. Massaloux, C. Lamblin, and J. Petit, "A silence compression scheme for use with g. 729 optimized for v. 70 digital simultaneous voice and data applications (recommendation g. 729 annex b)," *IEEE Commun. Mag*, vol. 35, no. 9, pp. 64–73, 1997.

[23] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3-4, pp. 271–287, 2004.

[24] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Sig. Proc. Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[25] H. Ghaemmaghami, D. Dean, S. Kalantari, S. Sridharan, and C. Fookes, "Complete-linkage clustering for voice activity detection in audio and visual speech," in *Interspeech*, 2015, pp. 2292–2296.

[26] D. A. Silva, J. A. Stuchi, R. P. V. Violato, and L. G. D. Cuozzo, "Exploring convolutional neural networks for voice activity detection," *Cognitive technologies*, pp. 37–47, 2017.

[27] F. Martinelli, G. Dellaferrera, P. Mainar, and M. Cernak, "Spiking neural networks trained with backpropagation for low power neuromorphic implementation of voice activity detection," in *ICASSP*. IEEE, 2020, pp. 8544–8548.

[28] M. Yang, C.-H. Yeh, Y. Zhou, J. P. Cerqueira, A. A. Lazar, and M. Seok, "A 1μw voice activity detector using analog feature extraction and digital deep neural network," in *ISSCC*. IEEE, 2018, pp. 346–348.

[29] G. Meoni, L. Pilato, and L. Fanucci, "A low power voice activity detector for portable applications," in *14th conference on PRIME*. IEEE, 2018, pp. 41–44.

[30] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, 2018.