

Voice Activity Detection (VAD) Using a Single-Layer Spiking Neural Network (SNN) Based on the Spike Response Model (SRM)

Asmaa Ourdighi ¹ *, Abdelkader Benyettou ²

¹ Computer Science Department, Faculty of Mathematics and Computer Science, University of Science and Technology of Oran Mohamed-Boudiaf, B.P. 1505 EL M'NAOUER 31000 Oran, Algeria, asmaa.ourdighi@univ-usto.dz

² University Center of Relizane, Algeria, a_Benyettou@yahoo.fr

Received: 17/11/2024

Accepted: 11/12/2024

Published: 19/12/2024

Abstract:

Speech is a primary medium of human communication, and automatic speech recognition (ASR) systems aim to transcribe spoken language into text. However, noise in audio signals can significantly degrade the performance of ASR systems. Voice Activity Detection (VAD) is a crucial pre-processing step that identifies speech segments within audio signals, helping to mitigate the impact of noise.

To improve the robustness of VAD systems, we propose a novel approach based on spiking neural networks (SNNs). SNNs are biologically inspired models that offer potential advantages in terms of energy efficiency and real-time processing. By leveraging the inherent capabilities of SNNs, we aim to develop a VAD system that is more resilient to noise and other adverse conditions.

Keywords: Voice Detection Activity (VAD), Spiking Neural Network (SNN), Spike Response Model (SRM), SpikeProp, Noisy environment.

* Dr. OURDIGHI Asmaa.

1. INTRODUCTION

Voice Activity Detection (VAD) is a fundamental component in numerous speech and audio applications, including speech coding, speech recognition, speech enhancement, and audio indexing. As a result, VAD is an essential initial phase in many speech processing systems, such as robust speech recognition and low-bandwidth speech compression systems (ITU-T Recommendation G.729-Annex B; ETSI EN 301 708 Recommendation). Essentially, a VAD algorithm aims to discern speech segments from non-speech regions within an audio signal.

A significant challenge in VAD is the presence of intense and non-stationary noise, which can significantly hinder the accurate detection of speech. Consequently, there has been a sustained effort to develop more robust VAD systems. However, despite decades of research, an optimal solution remains elusive.

Traditional VAD methods typically involve dividing the audio signal into short frames and extracting features such as energy, zero-crossing rate, and spectral coefficients. These features are then used to classify each frame as either speech or non-speech. While early methods relied on simple features, more recent approaches have employed complex features and statistical models to improve performance in noisy environments.

To enhance the robustness of VAD systems, researchers have explored various techniques, including the utilization of long-term temporal information to capture the broader context of speech signals, the exploitation of modulation features such as spectro-temporal modulation (STM) and amplitude modulation spectrogram (AMS) to improve detection in challenging conditions, the combination of complementary features to achieve better overall performance, the employment of neural networks like Radial Basis Function (RBF) networks and deep learning models to learn complex patterns in the

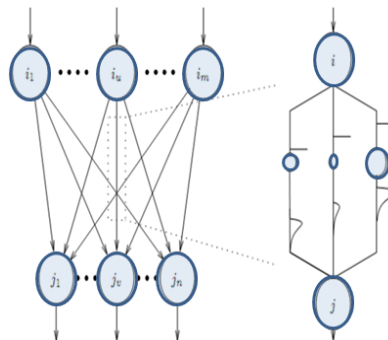
data, and the utilization of multi-microphone techniques to improve signal-to-noise ratio and localize the speech source.

In this work, we propose a novel approach to VAD using a single-layer Spiking Neural Network (SNN) based on the Spike Response Model (SRM). SNNs offer several advantages, including energy efficiency and the ability to process temporal information effectively. By training the SNN using supervised learning with the SpikeProp algorithm, we aim to develop a VAD system that is more robust to noise and better suited for real-time applications.

2. Proposed SNN model

Spiking Neural Networks (SNNs) stand out from traditional neural networks due to their diverse neuron models and learning algorithms, both supervised and unsupervised. The behavior of these algorithms is influenced by the network's architecture and the specific spiking model employed. Fundamentally, learning in SNNs involves the modification of synaptic weights, although some algorithms explore alternative mechanisms, such as adjusting connection delays (Maass, 1999).

Fig.1. Effects of selecting different switching under dynamic condition



For our approach, we adopt a spike response model with temporal delays, as proposed by (Gerstner, 1996) and further explored by (Bohte,

2002),. This model enables neurons to integrate input spikes over time, capturing temporal dependencies in the input signal. Each neuron is connected to a set of pre-synaptic neurons with varying delays within a specific range, allowing for the exploration of complex temporal dynamics

The refractory function η and the potential accumulation function ε , as defined in Eq. (1), govern the dynamics of the membrane potential $u_j(t)$ of neuron j . These functions regulate the neuron's response to incoming spikes and its ability to fire.

$$u_j(t) = \sum_{t_i^{(f)} \in F_j} \eta(t - t_i^{(f)}) - \sum_{i \in I} \sum_{t_i^{(g)} \in F_i} \sum_{k=1}^l w_{ji}^k \varepsilon(t - t_i^{(g)} - d^k) \quad (1)$$

The synaptic weight w_{ji}^k determines the strength of the connection between the pre-synaptic neuron i and the post-synaptic neuron j and denotes the weight associated with synaptic terminal k as a postsynaptic potential (PSP) of standard height with delay d^k ..

In Eq. (2), the post-synaptic potential (PSP) is modeled as the difference of two exponential functions, with time constants τ_m and τ_s controlling the rise and decay times, respectively. This model allows for a biologically realistic representation of synaptic transmission, where the timing of neural events plays a crucial role in information processing.

$$\varepsilon(s) = \left[\exp\left(-\frac{s}{\tau_m}\right) - \exp\left(-\frac{s}{\tau_s}\right) \right] H(s) \quad (2)$$

Where $H(s)$ denotes the heavy-side step function: $H(s)=0$ for $s \leq 0$ and $H(s)=1$ for $s > 0$.

After a neuron fires, it enters a refractory period, during which it is unable to fire again. This refractory period is modeled by Eq. (3). The τ_r parameter is another time-constant.

$$\eta(s) = -\vartheta \exp\left(-\frac{s}{\tau_r}\right) H(s) \quad (3)$$

A learning rule proposed by (Booij, 2005) is used to adjust the synaptic weights of our single-layer SNN. This rule minimizes the error between the desired and actual spike times of the output neurons, using a gradient descent approach. To calculate the error, only the first spike time of each output neuron is considered. The network error is defined as the sum of squared errors between the desired and actual first spike times of the output neurons.

$$E_{net} = \frac{1}{2} \sum_{j \in J} \left(t_j^1 - \hat{t}_j^1 \right)^2 \quad (4)$$

Where $\hat{t}_j^1$ denotes the desired first spike time and t_j^1 represented the calculated first spike time.

The parameters adjusted to reduce this error are the synaptic weights w_{ji}^k . The synaptic weight adaptation is calculated as follows in Eq. (5):

$$\Delta w_{ji}^k = -\eta \frac{\sum_{t_i^{(g)} \in F_i} \varepsilon(t_j^{(1)} - t_i^{(g)} - d^k)}{\sum_{i,k} \sum_{t_i^{(g)} \in F_i} w_{ji}^k \varepsilon'(t_j^{(1)} - t_i^{(g)} - d^k)} \left(t_j^{(1)} - \hat{t}_j^{(1)} \right) \quad (5)$$

Where η denotes the learning rate.

3. Experimental study

To conduct our experiments effectively, a crucial preprocessing phase is required to extract parameters that are well-suited for the Sequential Neural Network (SNN), ensuring accurate voice activity detection. This phase prepares both the 30 clean speech signals and their noise-corrupted counterparts, converting them into a simplified feature format. This streamlined representation will facilitate the integration of the data into the SNN, enhancing its learning process and ability to generalize across varying conditions.

3.1 Dataset used

Our experiments utilized the NOIZEUS database, designed for comparing speech enhancement algorithms, accessible at: <http://ecs.utdallas.edu/loizou/speech/noizeus/>. It contains 30 IEEE sentences recorded by diverse speakers in a soundproof booth. Originally at 25 kHz, the recordings were downsampled to 8 kHz, resulting in 80 seconds of data with 5000 frames for VAD (frame length

and overlap of 256). Notably, the speech content occupies 61.28% of the data, exceeding but slightly lower than (Gorriz, 2010). To simulate real-world noise, each clean sentence was corrupted with eight different noises from the AURORA database (crowds, cars, etc.) at varying signal-to-noise ratios (SNR). We specifically focused on four environments (airport, street, car, restaurant) at three SNR levels (5 dB, 10 dB, 15 dB), leading to a total of 720 corrupted sentences.

3.2 Preprocessing and SNN Parameters

To ensure accurate voice activity detection, we preprocessed the 30 clean and noisy speech signals, extracting relevant acoustic features. These features, derived from techniques like Linear Predictive Coding (LPC), were used to train the Sequential Neural Network (SNN), facilitating efficient learning and generalization.

To train the SNN, we manually labeled each 10ms frame of the clean signals as either speech or non-speech. These labeled frames, along with the extracted features, served as the training data for the SNN.

Fig. 2. Representation of the manual segmentation of the "sp01.wav" signal under the "Airport" environment: Clean (blue), 15 dB (red), 10 dB (yellow), 5 dB (green), and 0 dB (black).

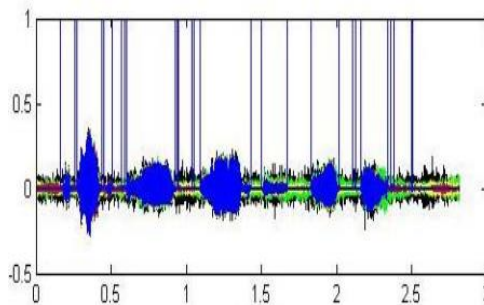


Figure 2 depicts the process of manually labeling clean speech signals. This labeling information is then extended to noisy signals at different SNR levels. For each frame, we extracted three features, including Linear Predictive Coding (LPC) coefficients, as described by (Kim, 2005).

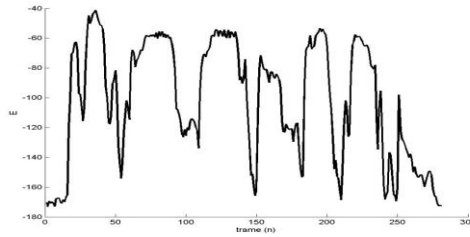
1) Short-time average power parameter (E)

The short-time average power (E) is a measure based on the logarithmic value of the zeroth-order autocorrelation function. It is computed using the following equation, where NNN represents the frame length in samples.

$$E = 10 \log \left[\frac{1}{N} \sum_{n=m-N+1}^m s(n)^2 \right] \quad (6)$$

This parameter tends to have low values during periods of silence or inactivity and higher values when speech is active. Fig. 3 shows how the short-time average power evolves over time in the clean audio signal "sp01.wav".

Fig. 3. Representation of the short-time average power parameter (E) from the clean signal "sp01.wav"



2) Zero-order most likelihood parameter (Z)

The zero-order likelihood (Z) parameter quantifies the spectral envelope's flatness, helping to distinguish between active and inactive voice regions. It generally takes higher values in regions containing speech. The parameter is calculated by taking the logarithm of the sum of squared Linear Predictive Coding (LPC) coefficients, as expressed in the Eq. (7).

$$Z = \log \sum_{i=0}^p a_i^2 \quad (7)$$

Where a_i are the LPC coefficients and p is the prediction order.

Fig.4. Representation of the zero-order likelihood parameter (Z) from the clean signal "sp01.wav".

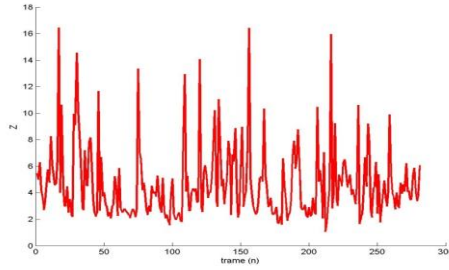


Fig.4 shows the evolution of zero-order likelihood parameter in the audio signal of our dataset.

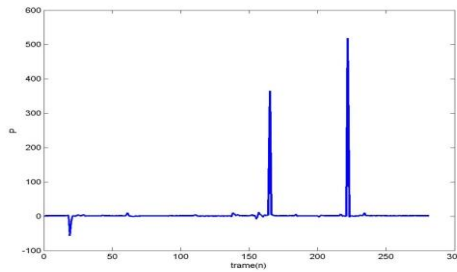
3) Pitch period difference parameter (P)

The Pitch Period Difference (P) parameter is derived from the maximum normalized autocorrelation and reflects the periodicity of the speech signal. It is used to characterize pitch stability, particularly useful in distinguishing between voiced and unvoiced segments of speech. The parameter is given by:

$$P = \max_{\tau} \left[\frac{\log \sum_{n=0}^{N-1} e(n)e(n-\tau)}{\log \sum_{n=0}^{N-1} e(n)e(n)} \right] \quad (8)$$

Where $20 \leq \tau \leq 160$ and (n) is the linear prediction error signal at time n .

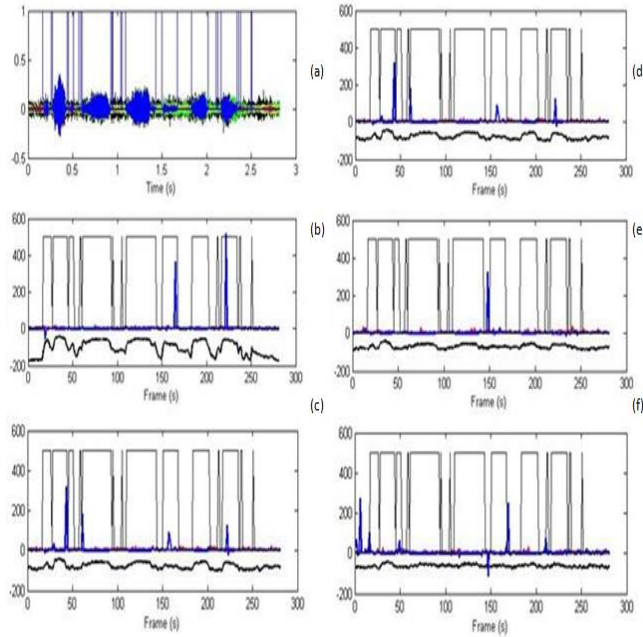
Fig. 5. Representation of the pitch period difference parameter (P) from the clean signal "sp01.wav".



The Fig.5 shows the values of Pitch period difference parameter in audio signal of our dataset.

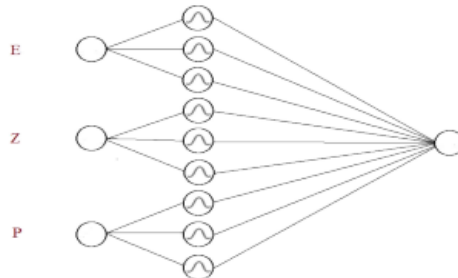
As shown in Figures 3, 4, and 5, the parameters E and Z exhibit high values during active speech segments, while parameter P remains relatively stable. In contrast, during periods of silence, these parameters show the opposite behavior. This trend is particularly evident when comparing these figures to Fig.2, which depicts the clean speech signal.

Fig. 6. Representation of the variations in calculated signal features under the Airport environment: Clean (b), 15 dB (c), 10 dB (d), 5 dB (e), and 0 dB (f).



These characteristics are particularly effective in high-noise environments. Fig.6 shows how these parameters vary across different Signal-to-Noise Ratios (SNRs): 0 dB, 5 dB, 10 dB, and 15 dB.

Fig. 7. Structure of the proposed single-layer SNN for performing VAD



The proposed Spiking Neural Network (SNN) for Voice Activity

Detection (VAD) is shown in Fig. 7. It uses a single-layer structure with Spike Response Model (SRM) units. In the input layer, population coding is applied, where each neuron is modeled with a Gaussian Receptive Field (GRF). The input features are encoded using six neurons, each with a GRF, spanning a 25 ms delay interval. The output layer consists of a single SRM neuron that makes the VAD decision. The system's response is based on phase encoding: a frame is considered active if the output neuron spikes within the [0, 15] ms phase window; otherwise, it is classified as inactive.

For the SRM parameters, the constants τ_m , τ_s and τ_r were initialized to 8, 4, and 8, respectively. The threshold for each neuron is initialized to 0.1. The number of sub-synapses for each connection varied between 6 and 10. As for the learning rate, after several trials, it was set to 10^{-3} .

3.3 Performance Matrices

To evaluate the VAD performance, we used the recognition rate (RR) and the Receiver Operating Characteristics (ROC) curve. The ROC curve, along with the Area Under the Curve (AUC) (Fawcett, 2003), helps visualize and compare the accuracy of the SNN across different noisy environments and varying Signal-to-Noise Ratio (SNR) levels (5 to 15 dB). A higher AUC, closer to 1, indicates better performance.

We also employed two additional evaluation metrics: the probability of correct decision in active zones (Pd) and the probability of correct decision in inactive zones (Nd), as defined by Eq. (8) and Eq. (9). These metrics provide a more detailed assessment of the VAD system's performance.

$$Pd = \frac{\text{Number of correct decisions in active zone(AZ)}}{\text{Total number of active zones}} \times 100 \quad (8)$$

$$Nd = \frac{\text{Number of correct decisions in inactive zones(NAZ)}}{\text{Total number of active zones}} \times 100 \quad (9)$$

The Pd (Probability of Detection) and Nd (Probability of False Detection) measures are key to assessing VAD performance. A higher Pd allows more voice to pass through, while a higher Nd reduces noise and non-voice signals. The closer both measures are to 100%, the better the VAD performance. However, Pd is more critical for audio quality: if Pd is above 90%, the voice message is clear and highly understandable. Between 80% and 90%, the message remains understandable, with minor interruptions. Below 80%, the message becomes harder to understand, and below 65%, it is largely incomprehensible.

3.4 Performance Measurement

After completing the preprocessing phase, we split the dataset evenly, allocating 50% for training and 50% for testing, to prepare the SNN for voice activity detection. During the generalization phase, we utilized 15 signals across four environments (Airport, Street, Car, and Restaurant) at SNR levels ranging from 5 to 15 dB.

The recognition rates of the SNN varied between 89% and 97.63%. Fig. 8, Fig. 9, Fig. 10 and Fig. 11 depict the AUC graphs for the four environments: Airport, Street, Car, and Restaurant.

Fig.8. AUC graphs of the proposed VAD based on SNN for the "Clean" signal, varying SNR in the "Airport" environment.

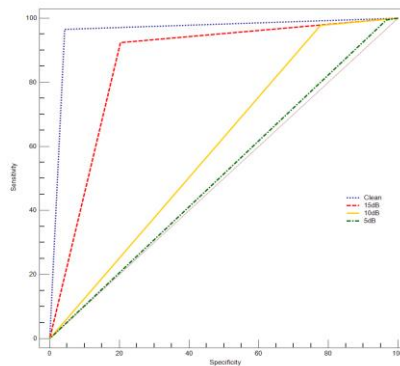


Fig.9. AUC graphs of the proposed VAD based on SNN for the "Clean" signal, varying SNR in the " Street " environment.

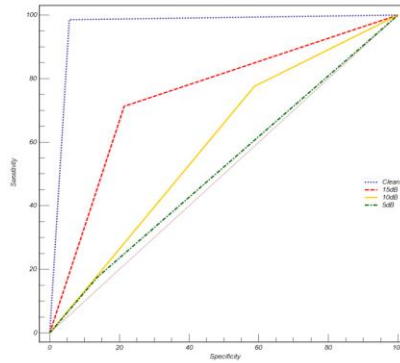


Fig.10. AUC graphs of the proposed VAD based on SNN for the "Clean" signal, varying SNR in the "Car" environment.

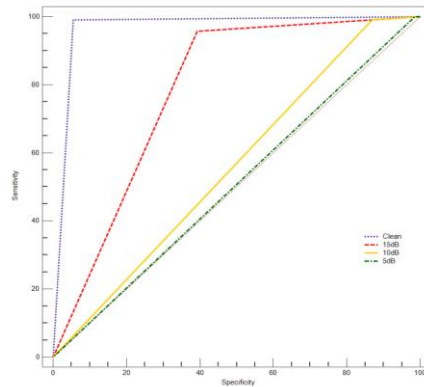


Fig.11. AUC graphs of the proposed VAD based on SNN for the "Clean" signal, varying SNR in the "Restaurant " environment

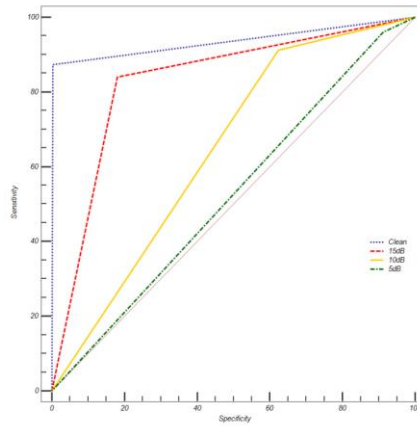


Table 1 presents the results of our proposed SNN-based VAD system. Our experiments reveal that the VAD's performance on clean signals directly impacts its robustness in noisy conditions. While the SNN excels on clean signals, its performance degrades significantly as the Signal-to-Noise Ratio (SNR) decreases below 10 dB. This is particularly evident in the Nd parameter, which indicates a higher rate of false alarms in noisy environments. Although this misclassification may not severely impact speech intelligibility after VAD processing, it highlights the limitations of the current system in extremely noisy conditions.

Table 1. The Results of the RR, AUC, Pd, and Nd of the proposed VAD based on SNN in the airport, street, car, and restaurant environments with a clean signal and SNR varying from 5 to 15 dB

		RR (%)	AUC (%)	Pd (%)	Nd (%)
Aéroport	Clean	96,27	0,961	96,59	95,66
	15dB	88,10	0,861	92,50	79,64
	10dB	71,95	0,600	97,82	22,26
	5dB	66,12	0,514	99,00	3,23
Street	Clean	97,15	0,965	98,62	94,34
	15dB	73,81	0,750	71,31	78,62
	10dB	65,16	0,594	77,61	41,21
	5dB	58,92	0,519	82,48	13,66

Voice Activity Detection (VAD) employing a single-layer spiking neural network based on the Spike Response Model (SRM)

Car	Clean	97,55	0,968	99,19	94,48
	15dB	83,75	0,782	95,75	60,69
	10dB	69,64	0,561	99,12	13,00
	5dB	66,24	0,507	100	1,39
Restaurant	Clean	91,62	0,936	87,41	99,70
	15dB	83,35	0,830	84,09	81,92
	10dB	72,83	0,644	91,20	37,54
	5dB	66,17	0,524	96,06	8,74

4. CONCLUSION

In this study, we introduced a single-layer Spiking Neural Network (SNN) for Voice Activity Detection (VAD). Despite its simplicity, the SNN showed promising results in terms of both training and testing efficiency. However, when compared to state-of-the-art VAD methods, the performance of our SNN, particularly in noisy environments, was somewhat limited. While the SNN performed well under controlled experimental conditions, its robustness in challenging, real-world scenarios needs improvement. Future research should focus on enhancing the SNN's ability to handle noisy conditions and improve its generalization to ensure better performance in practical applications.

5. Bibliography

1. Theses:

Chiodi. R. (Septembre 2010) Détection D'activité Vocale Basée sur la transformée enondelettes. Mémoire de maîtrise en Génie Électrique, Université du Québec à Trois-Rivières.

2. Journal article :

Gorriz, J.M., Ramirez, J. Lang, E.W., Puntonet, C.G., I. Turias, (2010) Improved Likelihood Ratio Test Based Voice Activity Detector Applied to Speech Recognition”, Speech Communication, Vol.52, pp.664–677.

Gerstner, W., Kempter, R. , Van Hemmen, L.J. Wagner, H. (1996) A neuronal learning rule for sub-millisecond temporal coding”, *Nature*, vol.384, pp.76–78.

Bohte, S.M. , Kok, J.N., Poutre, H. La (2002) Error-backpropagation in temporally encoded networks of spiking neurons, *Neurocomputing*, Vol. 48, pp.17–37.

Booij, O., Tat Nguyen, H. (2005) A gradient descent rule for spiking neurons emitting multiple spikes, *Inf. Process. Lett*, Vol.95, No.6, pp.552-558.

3. Seminar article:

Kim, H.I., Park, K.S. (2005) “Voice Activity Detection Algorithm Based on Radial Basis Function Network”, *IEICE Trans. Communication*, Vol. E88-B, No. 4, pp. 1656-1657.

Maass, W. , Schnitger, G., Sontag, E. (1991) On the computational power of sigmoid versus boolean threshold circuits, *Proc. of the 32nd Annual IEEE Symposium on Foundations of Computer Science*, p. 767-776.