COMPARATIVE ANALYSIS OF VOICE ACTIVITY DETECTION METHODS IN SPEECH SIGNAL PROCESSING

Kamoliddin Shukurov
Umidjon Khasanov
Shokhrukhmirzo Kholdorov
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi Email: umidjon0923@gmail.com

Abstract: *This paper analyzes various Voice Activity Detection (VAD) methods, which play a crucial role in speech signal processing. The main objective of VAD algorithms is to distinguish speech segments from silence and background noise. The paper discusses different approaches, including energy-based methods, spectral feature-based algorithms, statistical modeling techniques, and modern machine learning models, particularly deep neural networks. The advantages and limitations of each method are addressed, with a special emphasis on their practical applicability. In our experiments, since signal filtering and noise suppression are applied prior to VAD, energy-based approaches demonstrated high effectiveness and reliability.*

Keywords: *speech signal, VAD, energy-based approach, spectral features, statistical modeling, deep learning, filtering.*

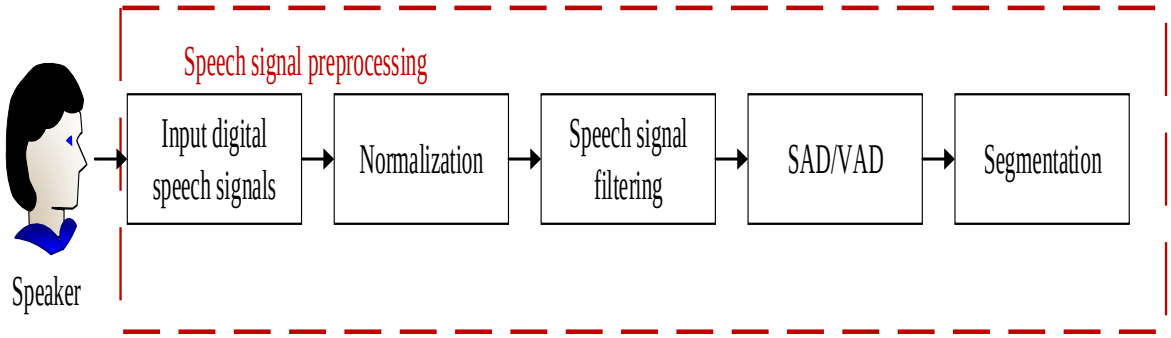
INTRODUCTION

In speech signal processing systems, one of the fundamental initial stages is Voice Activity Detection (VAD), which has a direct impact on the overall performance of the system. The primary task of VAD is to determine which segments of an acoustic signal contain human speech and, conversely, which parts consist only of noise or silence [1]. VAD is an essential component in speech recognition, compression codecs (e.g., discontinuous transmission during silence), and many other interactive voice-based systems.

However, VAD algorithms encounter serious challenges under various environmental conditions, particularly when the noise level is high.

Although many methods have been proposed in recent years to improve performance in low signal-to-noise ratio (SNR) conditions, the robustness of VAD is still not fully ensured at very low SNR values or in the presence of previously unseen noise types.

Consequently, developing and enhancing VAD methods that can operate reliably under different conditions remains a highly relevant scientific and technical task [1,2].



MAIN PART

Energy-Based and Zero-Crossing Rate Methods. The simplest and historically earliest approaches to VAD rely on short-term signal energy and the Zero-Crossing Rate (ZCR). In these methods, the input signal is divided into short frames, and the short-term energy of each frame is calculated. If the energy exceeds a predefined threshold, the frame is considered as containing speech; otherwise, it is classified as silence or noise.

Similarly, ZCR counts the number of times the signal crosses the zero axis within a frame, which helps to distinguish voiced and unvoiced segments. In practice, short-term energy and ZCR have been widely used to detect the start and end points of speech segments [1].

$$E(n) = \sum_{t=1}^{T} |x(t + nH) \cdot \omega(t)|^2 \quad (1)$$

where:

$E(n)$ – energy of the n-th frame,

$x(t)$ – signal,

$\omega(t)$ – windowing function.

These methods are computationally lightweight and fast, which is why they have been commonly adopted in applications such as voice command detection or silence compression (DTX) in telecommunication systems. Energy-based approaches perform well in clean environments but are highly sensitive to noise: when the background noise level unexpectedly increases, a simple threshold may misclassify noise as speech.

Spectral Feature-Based Methods. To improve the robustness of VAD against noise, approaches based on spectral features have been proposed. These methods analyze frequency-domain characteristics of the signal. For instance, spectral entropy measures the level of disorder in the spectrum. Experimental results show that spectral entropy values differ significantly between speech and noise-only segments. Studies have demonstrated that entropy-based methods can outperform simple energy-based approaches, particularly in scenarios where background noise varies over time [2].

$$F_{spec}(x_n) = -\sum_{f=1}^{F} p(f,n) \log p(f,n), \quad p(f,n) = \frac{|X(f,n)|^2}{\sum_{f=1}^{F} |X(f,n)|^2} \quad (2)$$

$$VAD(n) = \begin{cases} 1, & F_{spec}(x_n) < \theta \quad \text{(speech present)} \\ 0, & F_{spec}(x_n) > \theta \quad \text{(no speech)} \end{cases} \quad (3)$$

where:

$F_{spec}(x_n)$ – spectral feature function,

θ – threshold.

Other approaches in this category divide the signal into multiple subbands and assess the presence of speech within each band.

Statistical Model-Based Methods. To further refine VAD performance, statistical modeling and classical machine learning techniques have been introduced [3,5]. In statistical models, separate probability distributions are assumed for speech and non-speech states. Typically, Gaussian distributions are employed:

- H_0: no speech (only noise)
- H_1: speech+noise

For a given observation, the decision is made using the Likelihood Ratio Test (LRT):

$$\Delta(x) = \frac{p(x|H_1)}{p(x|H_0)} \quad (4)$$

$$VAD(n) = \begin{cases} 1, & \Delta(x) \geq \theta_{LRT} \\ 0, & \Delta(x) < \theta_{LRT} \end{cases} \quad (5)$$

These methods have been shown to provide more reliable performance than simple thresholding, especially in moderately noisy conditions [3].

Machine Learning and Deep Neural Network-Based Methods. In recent years, deep learning-based methods have emerged as a new stage in VAD research. Deep neural network (DNN) models achieve high accuracy in challenging environments, including very noisy conditions. One of their main advantages is the ability to learn discriminative features automatically when trained on large datasets [7,8].

The output of a neural model is generally interpreted as the probability of speech presence:

$$P(speech|x_n) = f_\theta(x_n) \quad (6)$$

$$VAD(n) = \begin{cases} 1, & P(nutq|x_n) \geq \theta_{ML} \\ 0, & P(nutq|x_n) < \theta_{ML} \end{cases} \quad (7)$$

where f_θ denotes the machine learning model (e.g., SVM, CNN, RNN, LSTM).

Although these models achieve state-of-the-art performance, they require significant computational resources and large amounts of training data.

CONCLUSION

Different VAD methods – ranging from simple energy-based techniques to more advanced spectral, statistical, and deep learning approaches – each have their strengths and weaknesses. Spectral features provide robustness in noisy conditions, statistical models deliver reliability in medium-noise environments, while neural networks achieve the highest accuracy in complex real-world scenarios but at the cost of computational complexity.

In our study, since signal filtering and noise suppression are applied prior to VAD, the input signal is already cleaned. This significantly reduces the need for more complex spectral or statistical methods. Under these conditions, simple energy-based VAD is sufficient to provide high accuracy and reliability. Moreover, it offers the advantages of computational efficiency and suitability for real-time applications.

REFERENCES:

1. N. N. Lokhande, P. S. Vikhe, N. S. Nehe, “Voice Activity Detection Algorithm for Speech Recognition Applications,” Int. Conf. in Computational Intelligence (ICCIA), 2011

2. K.-Q. Wang, T.-L. Hou, C.-L. Chin, “Voice Activity Detection Using Spectral Entropy in Bark-Scale Wavelet Domain,” Oriental COCOSDA Conference, 2009

3. J. Sohn, N. S. Kim, W. Sung, "A Statistical Model-Based Voice Activity Detection," IEEE Signal Processing Letters, vol. 6, no. 1, pp. 1–3, 1999

4. R. M. Patil, C. M. Patil, "Unveiling the State-of-the-Art: A Comprehensive Survey on Voice Activity Detection Techniques," 2025

5. S. Rajanayagam, M. A. Ingrisch, P. Müller, et al., "Enhancing Voice Activity Detection for an Elderly-Centric Self-Learning Conversational Robot Partner in Noisy Environments," Proc. ICAIIT, Apr 2025

6. S. Li, Y. Li, T. Feng, J. Shi, P. Zhang, "Voice Activity Detection Using a Local-Global Attention Model," Applied Acoustics, vol. 195, p.108802, 2022

7. K. Tripathi, C. V. Kumar, P. Wasnik, "Attention Is Not Always the Answer: Optimizing Voice Activity Detection with Simple Feature Fusion," arXiv preprint arXiv:2306.00910, 2025

8. X.-L. Zhang, M. Xu, "AUC Optimization for Deep Learning-based Voice Activity Detection," Proc. Interspeech, 2022