Novel Turbo Cancellation ISI/ICI DSP Concepts for FBMC-OQAM Based MIMO 5G Digital Communication Multi-Carrier Systems

Hemant Subhash Badodekar
VTU Research Scholar
hsbadodekar@gmail.com

Rakesh Subhash Badodekar
VTU Research Scholar
rsbsit@gmail.com

Dr B.G, Nagaraja
VTU Research Supervisor

***Abstract*: This article provides an all-encompassing survey on voice activity detection (VAD) a critical component crucial to speech processing systems responsible for detecting speech presence within an audio signal. The presented survey covers fundamental aspects related to VAD including its importance, applications, and inherent challenges faced during implementation. Our exploration initiates with establishing a solid foundation concerning the basics of VAD encompassing features and techniques in detail. Additionally key issues encountered along with challenges faced when implementing this technology efficiently is addressed. Fur- thermoses, it also delves into evaluation metrics commonly utilized for assessing overall performance whilst providing a comprehensive overview of readily accessible VAD databases. Overall, this survey predominantly presents a clear comprehend- Sion of VAD, the encountered challenges and the utilized techniques designed to overcome them ultimately serving as an esteemed resource for both researchers and professionals functioning within the speech processing field.**

Index Terms: VAD; ZCR; CNN; Metrics; Databases.

I. INTRODUCTION

Voice activity detection (VAD) primarily focuses on detecting whether an audio signal contains speech or non-speech activity. It is imperative in various applications such as audio compression, speech recognition [1], speaker identification [2], etc. allowing for the identification of segments that consist of speech. The VAD algorithm effectively analyzes the audio signal based on predetermined set criteria to establish if it comprises any form of speech [3]. Assessment criteria may include features such as energy levels, spectral density, zero crossing rate along with more advanced methodologies employing machine learning algorithms. There exist two primary types of VAD methods: frame based, and stream based. While frame-based VAD processes each frame individually within the audio signal. Stream based VAD treats the entire audio stream as one complete entity.

Detecting speech in a noisy environment presents greater challenges compared to a clean environment [4]. The presence of background noise can interfere with the clarity of the speech signal making it difficult to distinguish between speech and noise [5]. Additionally certain sources of noise such as wind, traffic, or other environmental factors can be false for speech further complicating accurate detection. To address these difficult- ties VAD algorithms designed for noisy environments employ advanced techniques that consider the characteristics of both the noise and speech signals. These techniques may include:

Spectral analysis: By analyzing frequency content, these methods are able to separate the speech signal from the background noise [6]. Generally, the energy in high frequency ranges is higher in speech signals than in noise signals which are more evenly distributed across all frequencies.

Statistical modeling: Statistical properties of speech and noise signals are modeled individually by this approach and utilized to classify whether a signal is speech or non-speech [7].

Adaptive filtering: used in VAD algorithms to estimate and subtract the noise signal from the input signal and leaving behind only the desired speech signal [8].

Machine learning algorithms: These algorithms use a training set of labeled speech and non-speech data to learn the characteristics of speech and noise signals and classify them accurately [9].

Overall, VAD in a noisy environment is an active area of research, and new techniques are continually being developed to improve the accuracy and robustness of VAD algorithms. This article provides an overview of VAD technologies, covering some representative techniques from the 1980s until the present day. The focus is on recent techniques that represent a paradigm shift from traditional methods to machine learning techniques. The article is intended to serve as a concise introduction to the research questions and solutions for those interested in starting research in VAD. It may also be useful for speech scientists who want to stay up-to-date on current trends in the field. Basic familiarity with digital signal processing and pattern recognition is assumed. Section 2 provides fundamentals of VAD. Section 3 describes the different VAD techniques. Literature work are elaborated in Section 4. Section 5 is then devoted to the different VAD metrics. We discuss the databases for the VAD research in Section 6, followed by conclusions in Section 7.

II. FUNDAMENTALS

Figure 1 shows the components of VAD system. The different components of the VAD system are:

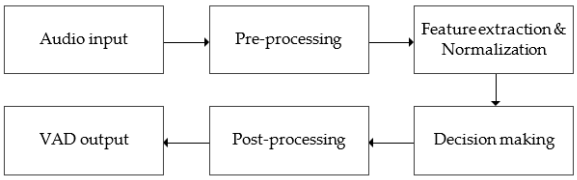

Fig. 1. Block diagram of VAD.

- Audio input: The audio input for a VAD system can either originate from a micro- phone or an audio file prior to processing by the VAD system.
- Pre-processing: It is common practice to preprocess the audio signal by removing additional noise sources. Filtering out components that are not related to human speech and equalizing its overall quality.
- Feature extraction & Normalization: The audio signal, which has undergone pre- processing is then examined to identify important characteristics such as energy, zero crossing rate (ZCR) and spectral content. These extracted characteristics are then normalized to account for variations in signal level and background noise.
- Decision making: The normalized features are subsequently inputted into a decision- making algorithm that categorizes the audio signal as either speech or non-speech. This algorithm can be based on statistical models, machine learning algorithms, or a combination of both.
- Post-processing: To refine the accuracy output of the decision-making algorithm may undergo further processing to eliminate short duration false positives or false negatives.
- Output: Ultimately. the VAD system produces a binary decision indicating whether there is speech present in the input audio signal. This output can be utilized by other speech processing applications like speech recognition or speaker identification.
- Traditional VAD techniques rely on crafted features and statistical models. However, recent advancements in machine learning have introduced data driven approaches like deep neural networks that can automatically extract relevant features from audio signals and make more precise decisions.

A. Selection of Features

- VAD systems heavily depend on extracting pertinent characteristics from audio signals. The selection of these features significantly impacts the effectiveness of the VAD system. Here are some common features used in VAD:
- Energy: Energy is commonly used as a feature in VAD and measures the power level of an audio signal by calculating the sum of squared amplitudes over a short time period.
- ZCR: The ZCR is the number of times the audio signal crosses the zero axis in a given time interval. It can be used as a feature in VAD to distinguish between speech and silence.
- Spectral features: Spectral features capture the frequency content of the audio signal. Common spectral features used in VAD include Mel-frequency cepstral coefficients (MFCCs) and spectral flux.
- Pitch: Pitch is the perceived fundamental frequency of a sound. It can be used as a feature in VAD to distinguish between speech and non-speech sounds, such as noise or music.
- Duration: The duration of the audio signal is also an important feature in VAD. Speech signals typically have longer durations compared to non-speech signals.
- Modulation: Modulation refers to the variations in the amplitude or frequency of the audio signal over time. Modulation features can be used in VAD to distinguish between speech and non-speech signals.

Table 1 shows the common features used in VAD.

TABLE I. COMMON FEATURES USED IN VAD.

Feature	Description
Energy	Measures the power level of an audio signal by calculating the sum of squared amplitudes over a short time period.
ZCR	Counts the number of times the audio signal crosses the zero axis in a given time interval, helping to distinguish between speech and silence.
Spectral Features	Capture the frequency content of the audio signal. Common examples include Mel-frequency cepstral coefficients (MFCCs) and spectral flux.
Pitch	Represents the perceived fundamental frequency of a sound, useful in distinguishing between speech and non- speech sounds like noise or music.
Duration	Speech signals generally have longer durations compared to non-speech signals, making this a useful feature in VAD.
Modulation	Refers to variations in amplitude or frequency over time, aiding in distinguishing between speech and non-speech signals.

B. Decision Making

There are several decision-making techniques used in VAD to determine whether speech is present in an audio signal or not. Some of the commonly used decision-making techniques in VAD are:

- Energy-based VAD: In this technique, the energy level of the audio signal is calculated and compared to a threshold value. If the energy level of the signal is above the threshold, it is considered as speech, and if it is below the threshold, it is considered as non-speech [10].
- ZCR VAD: The ZCR is the rate at which the signal changes sign. In this technique, the ZCR of the audio signal is calculated and compared to a threshold value. If the ZCR is above the threshold, it is considered as speech, and if it is below the threshold, it is considered as non-speech.
- Spectral-based VAD: In this technique, the spectral characteristics of the audio signal are analyzed to determine whether it contains speech or not. Various spectral- based features, such as Mel Frequency Cepstral Coefficients (MFCC), are calculated, and a

classifier is trained to differentiate between speech and non-speech segments.

- Hidden Markov model (HMM) VAD: In this technique, an HMM is used to model the speech and non-speech segments of the audio signal. The HMM is trained using a set of training data and used to classify the speech and non-speech segments of the audio signal [11].
- Neural network VAD: This method involves training a neural network to distinguish between speech and non-speech audio signal segments. The features utilized to train the neural network might be based on the energy, spectral, or time-frequency do- main properties of the audio signal. The neural network is taught using a collection of training data.

These decision-making techniques can be used individually or in combination to improve the accuracy of the VAD system. The selection of the appropriate technique depends on the application requirements and the characteristics of the audio signal being analyzed.

III. VAD TECHNIQUES

There are many different techniques that can be used for VAD, each with its own advantages and disadvantages. Here are few commonly used VAD techniques, along with the necessary equations.

A. Short-time Energy (STE)-Based VAD

This is a simple and widely used technique that uses the energy of the signal to detect speech. Let x(n) be the input signal and E(n) be the energy of the signal in frame n.

$$E(n) = \sum_{i=n-N+1}^{n} |x(i)|^2$$

where N is the window length over which the energy is calculated. The decision function is defined as:

$$D(n) = \begin{cases} 1; & \mathrm{E(n)} > \theta \\ 0; & \text{Otherwise} \end{cases}$$

where θ is a threshold value. The output of the energy-based VAD can be a binary decision indicating the presence or absence of speech, or it can be a continuous decision indicating the likelihood of speech being present in the signal.

B. ZCR Based VAD

ZCR based VAD is another commonly used technique for detecting speech in an audio signal. It involves calculating the rate at which the signal changes sign (i.e., crosses the zero axis) and comparing it to a threshold value to determine whether speech is present or not. Figure ?? shows the graphical representation of ZCR. The ZCR of the signal can be calculated using the following equation:

$$ZCR(n) = \frac{1}{2N} \sum_{i=n-N+1}^{n} |sign[x(i)] - sign[x(i-1)]|$$

The decision function is defined as:

$$D(n) = \begin{cases} 1; & \mathrm{Z(n)} > \theta \\ 0; & \text{Otherwise} \end{cases}$$

C. STE and ZCR Combined VAD

This technique combines the energy and ZCR of the signal to detect speech. The basic idea behind this technique is that speech signals have a higher energy and higher ZCR compared to non-speech signals. Therefore, by combining the energy and ZCR measures, it is possible to improve the accuracy of speech detection in a noisy environment. The STE is calculated as the average energy of the signal over a short time window, and the ZCR is calculated as the rate at which the signal crosses the zero-axis over the same window. The combined measure of the STE and ZCR can be calculated as follows:

$$C(n) = k \times \mathrm{STE}(n) + (1-k) \times \mathrm{ZCR}(n)$$

where C(n) is the combined measure of STE and ZCR at time index n, STE(n) is the short-time energy at time index n, ZCR(n) is the zero-crossing rate at time index n, and k is a weighting factor that determines the relative importance of STE and ZCR. The combined measure C(n) is compared to a threshold value to determine whether speech is present or not. If C(n) is above the threshold, it is considered as speech, and if it is below the threshold, it is considered as non-speech.

D. HMM based VAD

It is a statistical technique that models the speech and non-speech segments of an audio signal as different states of a Markov chain [12]. The HMM-based VAD uses a set of observed features, such as the frequency content, energy, or spectral envelope, to classify the signal into different states. The basic idea behind the HMM-based VAD is to model the probability distribution of the observed features of the speech and non-speech segments of the signal. This is done by training separate HMMs for the speech and non-speech states using a set of labeled training data. The observed features of the training data are used to estimate the parameters of the HMMs, such as the mean and variance of the probability distribution.

Let x(n) be the input signal, y(n) be the output decision, and Si be the ith state in the HMM. The HMM for speech can be represented as:

$$P(S_i|S_{i-1}) = A_{i-1,i}$$

$$P(x(n)|S_i) = B_i(x(n))$$

$$P(S_1) = \pi_1$$

The HMM for non-speech can be similarly represented. The decision function is defined as:

$$D(n) = \begin{cases} 1; & P(\text{speech}|x(n)) > P(\text{non-speech}|x(n)) \\ 0; & \text{Otherwise} \end{cases}$$

E. Gaussian mixture model (GMM) based VAD

It is a popular technique used to detect speech in noisy environments. In this method, speech and non-speech segments are modeled using separate GMMs [13]. The input signal x(n) is modeled as a mixture of K Gaussian distributions with weights wik, means μik, and covariances Σik, where i represents the speech or non-speech model and k represents

the k-th Gaussian component of the mixture model [14]. The likelihood of x(n) being generated from the ith GMM is given by:

$$p(x(n)|\lambda_i) = \sum_{k=1}^{K} w_{ik}\mathcal{N}(x(n)|\mu_{ik}, \Sigma_{ik})$$

where λi represents the GMM parameters for the i-th model and N (x(n)|μik, Σik) is the probability density function of a Gaussian distribution with mean μik and covariance Σik. The decision function for GMM-based VAD is defined as:

$$D(n) = \begin{cases} 1; & p(x(n)|\text{speech}) > p(x(n)|\text{non-speech}) \\ 0; & \text{Otherwise} \end{cases}$$

where $p(x(n)|\text{speech})$ and $p(x(n)|\text{non-speech})$ are the likelihoods of the input signal $x(n)$ being generated from the speech and non-speech GMMs, respectively. In the detection phase, the input signal is tested against the trained GMMs to detect the presence of speech. The performance of GMM-based VAD depends on the number of Gaussian components, the selection of the speech and non-speech models, and the choice of the threshold for the decision function.

F. Support vector machine (SVM) based VAD

SVMs are another popular technique used for VAD. In this method, an SVM classifier is trained to distinguish between speech and non-speech segments in the input signal. The input signal is first pre-processed to extract relevant features, such as MFCCs or linear prediction coefficients (LPCs). The feature vector for each frame of the input signal is then used as input to the SVM classifier. The SVM classifier is trained using a set of labeled training data [15]. The training data consists of feature vectors and their corresponding class labels (speech or non-speech). In the detection phase, the feature vector for each frame of the input signal is tested against the trained SVM classifier. The output of the SVM classifier is a continuous value between -1 and 1, which represents the distance of the feature vector from the hyperplane. A positive value indicates that the feature vector is more likely to be speech, while a negative value indicates that it is more likely to be non-speech.

The decision function for SVM-based VAD is defined as:

$$D(n) = \begin{cases} 1; & f(x(n)) > \theta \\ 0; & \text{Otherwise} \end{cases}$$

where $f(x(n))$ is the output of the SVM classifier for the input signal $x(n)$, and θ is a threshold value chosen to optimize the performance of the VAD. The SVM classifier is trained using the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\xi_i$$

subject to:

$$y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

G. Deep neural network DNN based VAD

DNNs have been shown to be effective for VAD in recent years [16]. DNN-based VAD involves training a neural network to classify each frame of the input signal as speech or non-speech. The input signal is first pre-processed to extract relevant features. The feature vector for each frame of the input signal is then used as input to the DNN. The DNN consists of multiple layers of nodes, with each layer transforming the input from the previous layer using a set of weights and biases. The output of the last layer is a single value between 0 and 1, which represents the probability that the input frame belongs to the speech class. The DNN is trained using a set of labeled training data. The training data consists of feature vectors and their corresponding class labels (speech or non-speech). The DNN learns to classify the input frames by adjusting the weights and biases in each layer to minimize the difference between its predicted outputs and the true class labels in the training data. In the detection phase, the feature vector for each frame of the input signal is tested against the trained DNN. The output of the DNN is a continuous value between 0 and 1, which represents the probability that the input frame belongs to the speech class. A threshold is then applied to the output to determine whether the input frame is speech or non-speech.

$$D(n) = \begin{cases} 1; & \hat{y}(n) > \theta \\ 0; & \text{Otherwise} \end{cases}$$

where $\hat{y}(n)$ is the output of the DNN for the input signal $x(n)$, and θ is a threshold value chosen to optimize the performance of the VAD. The DNN is typically trained using cross-entropy loss, which is defined as:

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)]$$

where θ represents the weights and biases of the DNN, N is the number of trainings samples, yi is the true class label of the ith training sample, and Yi is the predicted class label of the DNN for the ith training sample. The DNN can be trained using various optimization algorithms, such as stochastic gradient descent. In addition, various architectures can be used for the DNN, including convolutional neural networks, recurrent neural networks, and hybrid architectures.

H. Robust VAD (rVAD)

It is an unsupervised segment-based voice activity detection method that uses statistical modeling to detect speech segments in noisy environments. rVAD is

particularly useful in scenarios where labeled data is not available for training supervised VAD models [17]. rVAD first divides the input signal into short segments and calculates the energy and ZCR of each segment and then models the statistical distribution of the energy and ZCR using a mixture of Gaussians and a mixture of Laplacians, respectively. GMM for the energy distribution is defined as:

$$p_{\mathrm{GMM}}(x) = \sum_{i=1}^{K} w_i \mathcal{N}(x; \mu_i, \sigma_i^2)$$

where x is the energy of a segment, K is the number of mixture components, w_i is the weight of the i-th mixture component, μ_i is the mean of the i-th mixture component, and σ^2 is the variance of the i-th mixture component. $\mathcal{N}(x; \mu_i, \sigma^2)$ represents the Gaussian probability density function with mean μ_i and variance σ^2.

The Laplacian mixture model (LMM) for the ZCR distribution is defined as:

$$p_{\mathrm{LMM}}(x) = \sum_{i=1}^{K} w_i \mathrm{Lap}(x; \mu_i, b_i)$$

where x is the ZCR of a segment, K is the number of mixture components, wi is the weight of the i-th mixture component, μi is the mean of the i-th mixture component, and bi is the scale parameter of the i-th mixture component. Lap(x; μi, bi) represents the Laplace probability density function with mean μi and scale parameter bi. rVAD then calculates a speech likelihood score for each segment using the GMM and LMM models:

$$L(x) = \log p_{\mathrm{GMM}}(x) - \log p_{\mathrm{LMM}}(x)$$

The speech likelihood score is then used to detect speech segments in the input signal. A segment is classified as speech if its speech likelihood score is above a certain threshold, and non-speech otherwise. In summary,

Energy-based VAD:

- Simple and computationally efficient.
- Not robust to variations in noise environment or changes in signal characteristics.

STE and ZCR Combined VAD:

- More robust to noise and signal variations than energy-based or ZCR-based VAD alone.
- Combines energy and ZCR to compensate for weaknesses of each measure.
- Improves speech detection accuracy in noisy environments.

HMM-based VAD:

- Can be trained to model speech and non-speech characteristics for a specific application.
- Requires a large amount of labeled training data.
- Computationally expensive, making it less suitable for real-time applications.

GMM-based VAD:

- Models specific characteristics of speech and non-speech segments.
- More robust to variations in signal characteristics and noise environments.
- Requires a large amount of labeled training data for good performance.

SVM-based VAD:

- Effective in both clean and noisy environments.
- Compatible with various feature extraction techniques.
- May require more training data compared to other VAD techniques.

DNN-based VAD:

- Achieves state-of-the-art performance in various environments, including noisy and reverberant conditions.
- Requires a large amount of training data.
- High computational cost during training and testing.

rVAD:

- Robust performance in various noisy environments.
- May struggle to detect speech segments with low energy or high noise levels.
- Requires careful tuning of parameters for optimal performance.

Which VAD techniques one should use? It depends on various factors, including the specific application, the characteristics of the audio signals, and the available computational resources. For someone who would like to start research in VAD, we recommend to begin with the energy based and ZCR methods. Spectral-based methods, such as the MFCC and LPC, can provide better performance in noisy environments but may be computationally more demanding. Deep learning-based approaches, such as convolutional neural networks (CNNs) and recurrent neural networks may require more computational resources and larger amounts of training data [18]. To conclude, there is currently no universally recognized “best” VAD technique. Rather, the choice of technique typically involves a trade-off between maximizing accuracy and minimizing computational complexity and resource requirements.

IV. Literature

The study [19] evaluates a range of features, including energy-based features, ZCR, spectral centroid, MFCCs, and various combinations of these features. A dataset of clean and noisy speech signals was used to evaluate the performance of each feature, using several performance metrics. The results showed that a combination of energy-based features and MFCCs performed best in most cases, achieving high detection accuracy and low false alarm rates.

Furthermore, the study explored the impact of different window lengths on VAD performance, with the authors concluding that longer window lengths can improve VAD accuracy in noisy conditions. A new VAD algorithm that combines source and filter- based information to improve VAD performance in noisy environments was proposed in [20]. The study uses a DNN to extract features from speech signals and classifies them as speech or non-speech. The results showed that the proposed algorithm outperforms other VAD algorithms, achieving higher detection accuracy and lower false alarm rates. Furthermore, the study investigates the effect of different feature combinations on VAD performance, demonstrating that a combination of source-based and filter-based features improves VAD accuracy.

The work in [17] proposes a new unsupervised segment-based approach to VAD that is designed to be robust to noise and channel distortions. The proposed algorithm uses a combination of spectral subtraction and clustering techniques to identify speech segments in an audio signal. The paper also explores the impact of different types of noise and channel distortions on VAD performance, demonstrating that the rVAD algorithm is robust to various types of noise and channel distortions. In [21], a real-time VAD application designed for smartphone devices. The proposed app uses a CNN to classify audio signals as speech or non-speech, allowing for the real-time detection of speech activity. The paper also explores the impact of different types of noise and the effect of varying the CNN architecture on VAD performance. Additionally, the authors investigate the app's computational efficiency and show that the proposed approach is computationally efficient and suitable for real-time processing on smartphone devices.

A novel approach to VAD using time-delay neural networks (TDNNs) was presented in [22]. The proposed algorithm uses a TDNN to extract features from audio signals and classify them as speech or non-speech. The paper also explores the impact of different types of noise and varying the TDNN architecture on VAD performance. Additionally, the authors investigate the effect of different feature combinations on VAD performance, demonstrating that a combination of temporal and spectral features improves VAD ac- curacy. The study in [23] proposes a novel approach to optimize the area under the curve (AUC) of the receiver operating characteristic (ROC) curve for deep learning-based VAD. The authors argue that optimizing the AUC directly leads to better performance than optimizing traditional metrics such as accuracy, precision, and recall. The proposed approach uses a binary cross-entropy loss function to optimize the AUC during training of a CNN for VAD. The paper also explores the impact of varying the CNN architecture and the effect of different types of noise on VAD performance. Additionally, the authors investigate the robustness of the proposed approach to different training datasets, demonstrating that the proposed approach outperforms other VAD algorithms.

The study in [24] proposes a novel approach to VAD using dual attention mechanisms in both the time and frequency domains. The proposed approach uses a CNN with dual attention mechanisms in both the time and frequency domains to extract features from audio signals and classify them as speech or non-speech. The paper also explores the impact of varying the attention mechanisms in both the time and frequency domains on VAD performance. Additionally, the authors investigate the effect of different types of noise on VAD performance, demonstrating that the proposed approach is robust to different types of noise. Traditional VAD systems rely on digital signal processing algorithms that require high power consumption, which is a limitation for mobile devices and low-power embedded systems. To overcome this limitation, the authors in [25] pro- pose a low-power neuromorphic implementation of VAD using spiking neural networks (SNNs). SNNs are biologically inspired networks that mimic the behavior of neurons in the brain, and they are well-suited for low-power and real-time applications. The authors use a novel training algorithm based on spike-time-dependent backpropagation (STDB), to train the SNNs for VAD. STDB allows the network to learn the temporal dynamics of speech signals and adjust the synaptic weights accordingly.

A new approach to VAD in multi-speaker environments was proposed in [26]. The proposed method is based on polynomial eigenvalue decomposition (PEVD), which is used to extract the target speaker's voice from the mixed audio signal. The authors evaluate the performance of their system on a dataset of mixed speech signals with competing talkers and compare it to other state-of-the-art VAD systems. The results show that the proposed system achieves significantly better performance than other systems, with a detection error rate of 7.58%. The paper [27] presents a new approach to designing voice activity detection (VAD) systems using neural architecture search (NAS) techniques. The proposed NAS-VAD method automatically searches for the best VAD model architecture from a large space of possible architectures, optimizing both accuracy and efficiency. It was observed that the NAS-VAD approach requires significantly fewer parameters and computational resources than the other VAD systems, making it more efficient. Here are some of the overall conclusions that can be drawn from the VAD literature:

- Compared to traditional feature-based approaches, deep learning approaches have shown significant improvements in VAD performance.
- Several recent VAD studies have focused on developing systems that can handle competing talkers in multi-speaker environments.
- Attention mechanisms have emerged as a promising approach for VAD in multi- speaker environments, allowing the system to focus on the target speaker's voice while filtering out interfering sounds.
- A wide range of model architectures can be automatically selected from a large space using NAS techniques, which have shown significant promise for designing accurate and efficient VAD systems.
- Different VAD systems perform differently depending on the operating conditions, such as noise levels and microphone configurations, and there is a need to adapt VAD systems to these conditions.
- VAD systems can be integrated with other speech processing components, such as automatic speech recognition (ASR), speaker recognition, and speech enhancement, to improve overall system performance.

Table 2 shows the comparative analysis of the VAD techniques discussed in this work. In recent years, VAD literature has demonstrated significant progress, with new approaches being proposed to improve its accuracy and robustness. These methods have the potential to enable the development of more accurate and efficient speech processing systems for various applications. Research still needs to be done in this area, particularly in developing VAD systems that are robust to a variety of operating conditions, including different noises and microphone configurations. Nevertheless, recent advancements in deep learning, attention mechanisms, and NAS techniques have shown great promise in addressing these challenges.

V. VAD METRICS

VAD metrics are used to evaluate the performance of a VAD algorithm in detecting speech and non-speech segments of an audio signal [28]. Some commonly used VAD metrics are:

Detection accuracy (DA): The proportion of correctly detected speech and non- speech segments.

$$DA = \frac{S_P + N_P}{(S_P + S_F + N_P + N_F)}$$

where SP is the count of correctly detected speech segments, SF is the count of non-speech segments incorrectly detected as speech, NP is the count of correctly detected non-speech segments, and N_F is the count of speech segments incorrectly detected as non-speech.

False alarm rate (FAR): The proportion of non-speech segments that are incorrectly detected as speech.

$$FAR = \frac{S_F}{(S_F + N_P)}$$

Miss rate (MR): The proportion of speech segments that are incorrectly detected as non-speech.

$$MR = \frac{N_F}{(N_F + S_P)}$$

Precision: The proportion of detected speech segments that are actually speech.

$$\text{Precision} = \frac{S_P}{(S_P + S_F)}$$

Recall: The proportion of actual speech segments that are correctly detected.

$$\text{Recall} = \frac{S_P}{(S_P + N_F)}$$

F1-score: The harmonic mean of precision and recall.

$$\text{F1} - \text{score} = 2 \times \frac{\text{precision} \times \text{recall}}{(\text{precision} + \text{recall})}$$

Error rate (ER): The proportion of speech frames that are misclassified.

$$ER = \frac{(S_F + N_F)}{(S_P + S_F + N_P + N_F)}$$

Segmental signal-to-noise ratio (SSNR): The ratio of speech energy to noise energy for correctly detected speech segments [29].

$$SSNR = 10 \times \log_{10}\left(\frac{E_t}{E_b}\right)$$

TABLE II. COMPARATIVE ANALYSIS OF VAD TECHNIQUES

Reference	Methodology	Key Findings	Limitations
[19]	Evaluates various features (energy, ZCR, spectral centroid, MFCCs)	Combination of energy and MFCCs gives best performance; longer windows improve accuracy in noise	Limited to feature-based methods
[20]	DNN-based feature extraction and classification	Outperforms traditional VAD methods; source and filter-based features im- prove accuracy	Requires training data and computational resources
[17]	Unsupervised segment-based VAD using spectral sub- traction and clustering	Robust to noise and channel distortions	May not generalize well to all noise types
[21]	CNN-based real-time VAD application for smartphones	Computationally efficient and suitable for real-time processing	Performance de- pends on CNN architecture and noise conditions
[22]	TDNN-based feature extraction and classification	Combination of temporal and spectral features im- proves accuracy	Requires optimization ofTDNN architecture
[23]	AUC-optimized deep learning-based VAD	Optimizing AUC improves Performance over traditional metrics	Requires large datasets for training robust -ness
[24]	CNN with dual attention mechanisms intime and frequency domains	Robust to different types of noise	Attention mechanism tuning needed for optimal results
[25]	Low-power neuromorphic VAD using spiking neural networks (SNNs)	Energy-efficient for embedded systems; STDB-based training adapts well	Complexity in trainingand implementation
[26]	PEVD-based target speaker extraction in multi-speaker environments	Significant improvement in detection error rate (7.58%)	Requires high computational power
[27]	Neural Architecture Search (NAS)-based VAD optimization	Finds optimal VAD model with fewer parameters and lower computational cost	NAS process May be time- consuming

where E_t represents the energy of the speech signal, and E_b denotes the energy of the background noise.

Segmental signal-to-interference ratio (SSIR): The ratio of speech energy to interference energy for correctly detected speech segments.

$$SSIR = 10 \times \log_{10}\left(\frac{E_t}{E_i}\right)$$

where E_i represents the energy of the interference signal.

Equal error rate (EER): The rate at which the system incorrectly identifies speech segments as non-speech and non-speech segments as speech [30]. The lower the EER, the more accurate the VAD system.

$$EER = \frac{F_A + F_R}{2}$$

where F_A represents the false acceptance rate, and F_R represents the false rejection rate.

Minimum classification error (MCE): The sum of the false positives and false negatives divided by the total number of frames.

$$MCE = \frac{S_F + N_F}{(S_P + S_F + N_P + N_F)}$$

Frame-based accuracy (FA): The proportion of correctly classified speech frames.

$$FA = \frac{S_P + N_P}{(S_P + S_F + N_P + N_F)}$$

Intersection over union (IoU): The ratio of the intersection between the detected speech and the ground truth speech to their union.

$$IoU = \frac{S_P}{(S_P + S_F + N_F)}$$

In general, VAD metrics should be chosen in accordance with the requirements of the VAD system and the application. A good VAD system should achieve high values for DA, precision, recall, F1 score, SSNR, SSIR, AUC and low values for FAR, MR, MCE and IOU. A VAD system's specific requirements and application determine what metrics to use to measure its performance. Overall, a comprehensive analysis of these metrics can help researchers and developers improve VAD algorithms and optimize their performance for various speech processing applications. In order to assess a VAD system comprehensively, more than one metric should be used because no single metric can fully capture its performance.

VI. DATABASE FOR THE VAD RESEARCH

These databases provide a diverse range of audio recordings that can be used to evaluate the performance of different VAD techniques under various conditions, making them valuable resources for VAD research.

- TIMIT: This is a widely used database for speech research, containing recordings of sentences spoken by 630 speakers from eight major dialects of American English [31].
- CHiME: This is a dataset for speech recognition in noisy environments, containing audio recordings made in different environments with multiple speakers [32].
- VoxCeleb: This dataset contains audio recordings of speakers from various back- grounds and occupations, making it useful for speaker identification and VAD re- search [33].
- MUSAN: The Multi-lingual Speaker Identification and Verification (MUSAN) corpus is a collection of audio recordings in various languages and acoustic conditions, making it useful for developing VAD techniques that are robust to different lan- gauges and environments [34].
- DEMAND: The Demand dataset contains audio recordings made in various environments, including residential, urban, and office spaces, making it useful for studying VAD in real-world scenarios [35].
- Libri Speech: This is a dataset of audio recordings of speech from audiobooks, containing over 1000 hours of audio in English [36].
- Common Voice: This is a large dataset of audio recordings in multiple languages, contributed by people from around the world, making it useful for developing VAD techniques that are applicable in different linguistic and cultural contexts [37].
- NOIZEUS: The NOIZEUS database is a collection of audio recordings that was designed for speech enhancement and VAD research. The database consists of 60 audio files, each containing a mixture of speech and noise, with different types of noise [38] [39].
- MOSCOW: A collection of Russian speech recordings may be found in the MOSCOW database. It contains recordings of various speech patterns, such as read speech, spontaneous speech, and speech in noisy environments [40]. It is frequently used to assess how well speech coding systems function in various speech and noise scenarios.

Speech databases for VAD task is shown in Table 3. It is important to consider the specific research questions and study goals when selecting a VAD database. Different databases offer a range of recordings, including speech signals, in languages, background noise in various environments and emotionally expressive speech. Hence researchers need to evaluate which database suits their research requirements. It's also crucial to consider the size of the database since larger datasets provide training and testing data for VAD algorithms. Additionally, one should take into account the availability and accessibility of the corpus as certain databases may require permission or have usage restrictions. Overall selecting a database is vital for ensuring valid VAD research outcomes.

TABLE III. COMPARATIVE ANALYSIS OF SPEECH DATABASES.

Database	Language	Speakers	Text Content	Environment	Open Access	f_S (Hz)
CHiME	English	Multiple	Scripted Sentences	Realistic	No	16,000
Common Voice	Multiple	Multiple	User-contributed	Varied	Yes	48,000
DEMAND	Rennes (France)	Multiple	N/A	Realistic	Yes	16,000
LibriSpeech	Multiple	Multiple	Read Sentences	Controlled	Yes	16,000
MOSCOW	Russian	26	Varies	Varied	Not specified	44,100
MUSAN	Various	Single	Read speech	Varied	Yes	44,100
NIST	English	Multiple	Scripted Sentences	Controlled	Yes	16,000
NOIZEUS	English	06	clean speech	Varied	Yes	48,000
TIMIT	English	Multiple	Phonetically Rich	Controlled	No	16,000
VoxCeleb	Various	Multiple	N/A	Varied	Yes	Varies

VII. CONCLUSIONS

We have presented an overview of the classical and recent methods of VAD. The pay- per discusses features that help distinguish between speech and non-speech signals such as energy-based features, spectral based features, and statistical-based features. It also highlights challenges faced by researchers in developing VAD algorithms, such as striking a balance, between accuracy and computational complexity or ensuring robustness in noisy environments. Furthermore, the paper offers an overview of VAD techniques including model-based methods, rule-based methods, and machine learning based methods. Furthermore, the paper delves into the measures that can be employed to assess the effectiveness of algorithms used for VAD. Overall, this survey paper provides a comprehensive overview of the VAD literature and can serve as a valuable reference for researchers who are interested in developing or evaluating VAD algorithms. It highlights the need for further research in the field of VAD, particularly in developing more robust and efficient real-time VAD that can operate effectively in a variety of noisy environments.

DECLARATIONS

The authors have no conflict of interests on the manuscript.

REFERENCES

[1] Makhlouf, A., Lazli, L. and Bensaker, B., 2016. Evolutionary structure of hidden Markov models for audio-visual Arabic speech recognition. International Journal of Signal and Imaging Systems Engineering, 9(1), pp.55-66.

[2] Nagaraja, B.G. and Jayanna, H.S., 2016. Feature extraction and modelling tech- niques for multilingual speaker recognition: a review. International Journal of Signal and Imaging Systems Engineering, 9(2), pp.67-78.

[3] Ramirez, J., G´orriz, J.M. and Segura, J.C., 2007. Voice activity detection. funda- mentals and speech recognition system robustness. Robust speech recognition and understanding, 6(9), pp.1-22.

[4] Jainar, S.J., Sale, P.L. and Nagaraja, B.G., 2020. VAD, feature extraction and mod- elling techniques for speaker recognition: a review. International Journal of Signal and Imaging Systems Engineering, 12(1-2), pp.1-18.

[5] Heo, Y. and Lee, S., 2023. Supervised Contrastive Learning for Voice Activity De- tection. Electronics, 12(3), p.705.

[6] Graf, S., Herbig, T. and Buck, M., 2023. 13 Voice Activity Detection for In-Car Communication Systems. Towards Human-Vehicle Harmonization, 3, p.163.

[7] Zhu, Z. and Pei, K., A Robust Soft Voice Activity Detection Algorism Based on Multi-Feature Fusion Cosine Similarity at Low Signal-to-Noise Ratio. Available at SSRN 4345665.

[8] Bendoumia, R., Hassani, I. and Guessoum, A., 2023. Recursive adaptive filtering algorithms for sparse channel identification and acoustic noise reduction. Analog Integrated Circuits and Signal Processing, 114(1), pp.51-73.

[9] Chien, Y.R., Zhou, M., Peng, A., Zhu, N. and Torres-Sospedra, J., 2023. Signal Processing and Machine Learning for Smart Sensing Applications. Sensors, 23(3), p.1445.

[10] Pang, J., 2017, January. Spectrum energy-based voice activity detection. In 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 1-5). IEEE.

[11] Tan, Y.W., Liu, W.J., Jiang, W. and Zheng, H., 2014, July. Hybrid svm/hmm architectures for statistical model-based voice activity detection. In 2014 International Joint Conference on Neural Networks (IJCNN) (pp. 2875-2878). IEEE.

[12] Ouahabi, S.E., Atounti, M. and Bellouki, M., 2020. HMM-GMM based Amazigh speech recognition system. International Journal of Signal and Imaging Systems Engineering, 12(1-2), pp.47-53.

[13] Sholokhov, A., Sahidullah, M. and Kinnunen, T., 2018. Semi-supervised speech ac- tivity detection with an application to automatic speaker verification. Computer Speech & Language, 47, pp.132-156.

[14] Ouahabi, S.E., Atounti, M. and Bellouki, M., 2020. HMM-GMM based Amazigh speech recognition system. International Journal of Signal and Imaging Systems Engineering, 12(1-2), pp.47-53.

[15] Sandabad, S., Benba, A., Tahri, Y.S. and Hammouch, A., 2016. Novel extraction and tumour detection method using histogram study and SVM classification. Inter- national Journal of Signal and Imaging Systems Engineering, 9(4-5), pp.202-208.

[16] Fredj, I.B., Zouhir, Y. and Ouni, K., 2018. Fusion features for robust speaker identification. International Journal of Signal and Imaging Systems Engineering, 11(2), pp.65-72.

[17] Tan, Z.H. and Dehak, N., 2020. rVAD: An unsupervised segment-based robust voice activity detection method. Computer speech & language, 59, pp.1-21.

[18] Ashwini, B. and Yuvaraju, B.N., 2017. Application of machine learning approach in detection and classification of cars of an image. International Journal of Signal and Imaging Systems Engineering, 10(1-2), pp.8-13.

[19] Graf, S., Herbig, T., Buck, M. and Schmidt, G., 2015. Features for voice activity detection: a comparative analysis. EURASIP Journal on Advances in Signal Pro- cessing, 2015, pp.1-15.

[20] Drugman, T., Stylianou, Y., Kida, Y. and Akamine, M., 2015. Voice activity detection: Merging source and filter-based information. IEEE Signal Processing Letters, 23(2), pp.252-256.

[21] Sehgal, A. and Kehtarnavaz, N., 2018. A convolutional neural network smartphone app for real-time voice activity detection. IEEE Access, 6, pp.9017-9026.

[22] Bai, Y., Yi, J., Tao, J., Wen, Z. and Liu, B., 2019, November. Voice activity detection based on time-delay neural networks. In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 1173- 1178). IEEE.

[23] Zhang, X.L. and Xu, M., 2022. AUC optimization for deep learning-based voice activity detection. EURASIP Journal on Audio, Speech, and Music Processing, 2022(1), pp.1-12.

[24] Lee, J., Jung, Y. and Kim, H., 2020. Dual attention in time and frequency domain for voice activity detection. arXiv preprint arXiv:2003.12266.

[25] Martinelli, F., Dellaferrera, G., Mainar, P. and Cernak, M., 2020, May. Spiking neural networks trained with backpropagation for low power neuromorphic implementation of voice activity detection. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8544-8548). IEEE.

[26] Neo, V.W., Weiss, S., McKnight, S.W., Hogg, A.O. and Naylor, P.A., 2022, September. Polynomial eigenvalue decomposition-based target speaker voice activity detection in the presence of competing talkers. In 2022 International Workshop on Acoustic Signal Enhancement (IWAENC) (pp. 1-5). IEEE.

[27] Rho, D., Park, J. and Ko, J.H., 2022. Nas-vad: Neural architecture search for voice activity detection. arXiv preprint arXiv:2201.09032.

[28] Tong, S., Gu, H. and Yu, K., 2016, March. A comparative study of robustness of deep learning approaches for VAD. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5695-5699). IEEE.

[29] Nautsch, A., Bamberger, R. and Busch, C., 2016, September. Decision robustness of voice activity segmentation in unconstrained mobile speaker recognition environments. In 2016 International Conference of the Biometrics Special Interest Group (BIOSIG) (pp. 1-7). IEEE.

[30] Jayaprakash H., and Nagaraja B. G., 2020, A comparison of features for voice activity detection - a review and some experimental results, Vidyabharati International Interdisciplinary Research Journal, 9(2) (pp. 91-94)

[31] Zue, V., Seneff, S. and Glass, J., 1990. Speech database development at MIT: TIMIT and beyond. Speech communication, 9(4), pp.351-356.

[32] Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F. and Matassoni, M., 2013, December. The second 'CHiME' speech separation and recognition challenge: An overview of challenge systems and outcomes. In 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (pp. 162-167). IEEE.

[33] Nagrani, A., Chung, J.S. and Zisserman, A., 2017. VoxCeleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612.

[34] Snyder, D., Chen, G. and Povey, D., 2015. Musan: A music, speech, and noise corpus. arXiv preprint arXiv:1510.08484.

[35] Thiemann, J., Ito, N. and Vincent, E., 2013, June. The diverse environments multi- channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In Proceedings of Meetings on Acoustics ICA2013 (Vol. 19, No. 1, p. 035081). Acoustical Society of America.

[36] Panayotov, V., Chen, G., Povey, D. and Khudanpur, S., 2015, April. Librispeech: an as corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5206-5210). IEEE.

[37] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M. and Weber, G., 2019. Common voice: A massively- multilingual speech corpus. arXiv preprint arXiv:1912.06670.

[38] Hu, Y. and Loizou, P.C., 2007. Evaluation of objective quality measures for speech enhancement. IEEE Transactions on audio, speech, and language processing, 16(1), pp.229-238.

[39] Ma, J., Hu, Y. and Loizou, P.C., 2009. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. The Journal of the Acoustical Society of America, 125(5), pp.3387-3405.

[40] A.A. Zamyatnin, A.S. Borchikov, M.G. Vladimirov, O.L. Voronina, The EROP- Moscow oligopeptide database, Nucleic acids research, 34 (suppl 1) D261-D266 (2006)