# Noise Robust VAD Technique and its Performance Evaluation

ABHISHEK A S
*Department of Electronics and Communication Engineering*
*Vidyavardhaka College of Engineering*
Mysuru, India
abhishekas1501@gmail.com

TEJAS M B
*Department of Electronics and Communication Engineering*
*Vidyavardhaka College of Engineering*
Mysuru, India
mbtejas432@gmail.com

SRINIDHI S
*Department of Electronics and Communication Engineering*
*Vidyavardhaka College of Engineering*
Mysuru, India
ssrinidhi402@gmail.com

Dr. Nagaraja B G
*Department of Electronics and Communication Engineering*
*Vidyavardhaka College of Engineering*
Mysuru, India
nagarajabg@vvce.ac.in

SYED SHAZ MOHAMMED HUSSAINI
*Department of Electronics and Communication Engineering*
*Vidyavardhaka College of Engineering*
Mysuru, India
syedshaaz08@gmail.com

***Abstract*— In this real world of audio and speech signals, the ability to detect the voice activity and remove the background voice is important. This project presents that to remove noise robust Voice Activity Detection (VAD) technique that combines voice source and vocal tract system information using a zero-frequency filtering (ZFF) approach. As mentioned earlier ZFF technique is employed to combine and to compute a composite signal that encapsulate essential parameters such as fundamental frequency and formats, enabling robust VAD in time domain.This methodology offers a significant advantage in terms of computational efficiency compared to other methods, making it an attractive choice for real-time applications. By applying dynamic thresholding after spectral entropy-based weighting, this approach exhibits resilience or minimum across a range of Signal-to-Noise Ratios (SNRs), further enhancing its ability.This project provides comprehensive utility for audio segmentation, which offers the capability to process individual audio files or entire directories. It also includes the extraction of a composite signal, which is a critical component in enhancing voice quality and reducing background noise. With this, the project is open to use build further and use it for the big industries purpose and for communication purpose.**

**Keywords— Voice Activity Detection (VAD), Zero-Frequency Filtering (ZFF), Signal Noise Ratio (SNR).**

## I. INTRODUCTION (*HEADING 1*)

The real world has many of audio and signal receiving problems, it is very hard to detect the voice with the background noise or the background audio signals. The main intention of this project is to enhance the main audio signal and to remove the background noise as it helps to hear a clearer audio without any background noise. As there are many ways that are in market nowadays to remove the background noise, this is one of the different methods to provide user-friendly tools, and enhance voice quality. It's a journey towards a world where the human voice stands out amidst the noise, enabling seamless and intelligible interactions in a wide array of applications.

Voice Activity Detection is not just a feature but a linchpin in today's world of voice-controlled devices and seamless communication. It's the initial step to unlock the potential of voice data. Traditional VAD

often falters in noisy environments, underscoring the importance of a noise-robust approach.

The Noise Robust VAD Technique represents not just a technical innovation but a practical solution. It

strives to make VAD accessible, reliable, and robust, while enhancing voice quality and reducing

background interference. In an increasingly voice-centric world, this project aspires to elevate the clarity

of audio communication and set new standards in speech processing applications.

Voice Activity Detection (VAD): is a technology that identifies whether human speech is present or absent in an audio signal.

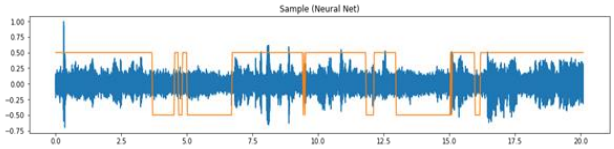

Figure1: Speech Recognition

VAD enables efficient processing and recognition of human speech in various applications, including telecommunications and voice-controlled systems.

VAD can face challenges in noisy environments, as high levels of background noise may lead to false positives (detecting non-speech as speech) or false negatives (missing speech in the presence of noise).

## II. NOISE ROBUST VAD TECHNIQUE

In general, noise-robust Voice Activity Detection (VAD) techniques aim to improve the accuracy of speech detection, especially in environments with background noise. Methods include analyzing signal energy, spectral features (e.g., spectral flux and entropy), statistical models (Gaussian Mixture Models, Hidden Markov Models), machine learning (supervised and deep learning), adaptive thresholding, multimodal approaches, and post-processing techniques like temporal smoothing. Environmental noise considerations and context-based decision fusion further enhance robustness. Effective strategies involve combining features and modalities, coupled with dynamic threshold adjustments based on Signal-to-Noise Ratio (SNR). These approaches collectively contribute to a more reliable VAD system in challenging acoustic conditions.

1.Signal Extraction:

- Clearer speech is also extracted.
- Within that noise are also extracted.

2.Multiplexer:

- Both clearer speech and noise are multiplied using multiplexer.

3.Enhancement mode:

- The noisy speech signal which is received from the multiplexer is enhanced using enhancement mode.

4.Voice Activity Detection:

- The signal is further proceeds to the VAD technique.

5.Voice Detection:

- The output signal is received with the noise removed speech signal which is more audible without any disturbance.

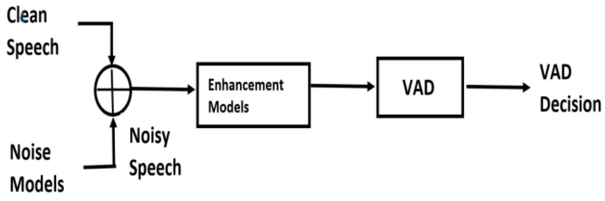

Figure 2: General Block Diagram Of NOISE ROBUST VAD TECHNIQUE

## II. LITERATURE REVIEW

M. Tejus Adiga and Rekha Bhandarkar [1] have introduced an enhanced Voice Activity Detection (VAD) method in their paper. This method integrates Adaptive Spectral Subtraction for noise cancellation with Single Frequency Filtering (SFF) to improve VAD performance, particularly during transitions between noise and speech. The approach includes a training phase to optimize filter coefficients for noise estimation and VAD filters. In the detection phase, the noise canceller estimates and suppresses noise in each frame before the SFF VAD filter determines whether the frame contains speech or noise. The system underwent evaluation on a MacBookPro with specific software and hardware configurations, demonstrating that the proposed solution enhances VAD accuracy, especially in challenging noise-to-speech transition scenarios.

H.M. Chang [2] have introduces a novel approach to noise spectral estimation tailored for highly non-stationary noise environments. It employs an entropy-based Voice Activity Detector (VAD) and incorporates recursive averaging with adaptive parameters to enhance accuracy in situations where noise characteristics undergo rapid changes. The methodology features an adaptive normalization parameter for VAD and a smoothing parameter adjusted based on speech presence probability for noise spectral estimation. The paper's contributions are noteworthy, showcasing improvements in segmental Signal-to-Noise Ratio (SNR) by 2.0-3.8dB and a reduction in noise spectral estimation error by 3.2-4.7dB, particularly in scenarios with swiftly changing noise types such as white and babble noise..

Kenji Nakayama; Shoya Higashi; Akihiro Hirano [3] conducted an investigation into Automatic Speech Recognition (ASR) performance across both landline and mobile networks. The methodology involved recording sessions using both types of phones, with participants reading words and phrases twice. The experiments evaluated ASR accuracy in speaker-dependent and independent scenarios, considering enrollment from landline, mobile, or both. The primary aim was to assess ASR performance degradation in cross-environment situations. The findings revealed significant performance drops when ASR systems trained in one environment were tested in another. However, the study demonstrated that cross-environment training could mitigate this degradation. A notable achievement of the research was the establishment of a benchmark for ASR performance degradation due to network variability, suggesting that training in multiple environments could improve accuracy

RUI WANG AND NAN LI [4] proposed method, the Masked Auditory VAD, leveraging a masked auditory encoder-based convolutional neural network (M-AECNN). The primary objective of this work is to improve the effectiveness of voice detection in noisy conditions. The method mimics human auditory processing using a gammachirp auditory filter bank (GAFB) and modulated speech signals to enhance VAD in challenging acoustic environments.By simulating the human ear's sound transmission to inner hair cells, the approach aims to enhance VAD performance in the presence of noise. The encoder component of the method reinforces robustness by emphasizing cleaner speech frequencies, leveraging the human ear's masking effect. This enhancement results in approximately 10.5% clearer voice detection. The study underscores the effectiveness of incorporating characteristics of the human auditory system into VAD systems. Future work is anticipated to extend this robust VAD approach to acoustic models.

WENPENG MU AND BINSHAN LIU [5] present the Adaptive Attention Span Transformer for Voice Activity Detection (AAT-VAD) in this paper, aiming to enhance processing efficiency and accuracy for extended audio recordings. The paper outlines the feature extraction process, incorporating pre-emphasis, windowed framing, and FFT and MEL frequency scale conversion. The AAT-VAD model is trained and validated using the TIMIT corpus, with the addition of noise to create a challenging dataset.Performance evaluation of the AAT-VAD model involves metrics such as F1-score, detection cost function, and average test time. Results indicate that AAT-VAD consistently outperforms existing technologies across various conditions. The primary objective of this research is to advance processing efficiency and accuracy in the context of long audio recordings or sequences. The paper seeks to bring attention to the field of Voice Activity Detection (VAD) and aims to inspire further research in this domain.

MOHAMMAD SHAHID AND CIGDOM BEYAN [6] introduces Speech-Visual Voice Activity Detection (S-VVAD), a system designed to detect speakers in videos by learning body motion associated with speech within a weakly merged segmentation framework. Operating directly on entire video frames without specific body part detection, S-VVAD addresses the VAD challenge, especially in challenging conditions where audio may be affected. Evaluating on panel discussions with varying challenges, the paper compares S-VVAD with state-of-the-art VVAD and multimodal VAD methods, showing comparable or slightly superior performance. The research contributes to the field by offering a solution to detect speakers in audio-visual recordings under diverse conditions.

D. Pravena and Govinda D [7] present the paper "Expressive Speech Analysis for Epoch Extraction Using Zero Frequency Filtering Approach." The paper addresses the challenge of epoch extraction in expressive speech, characterized by unpredictable pitch variations. The authors propose a method utilizing three Zero Frequency Resonators (ZFRs) and a trend removal process for short expressive speech segments. The approach aims to enhance epoch extraction accuracy and reduce spurious zero crossings by refining the Zero Frequency Filtered Signal (ZFFS).Evaluation on the German emotional speech database (EmoDb) demonstrates improved epoch extraction, measured through identification rate, miss rate, false alarm rate, and accuracy. The modified Zero Frequency Filtering (ZFF) method proves effective in capturing rapid pitch variations, particularly in emotions like anger, happiness, and fear, showcasing superior performance compared to conventional methods. The research contributes valuable insights into expressive speech analysis and epoch extraction techniques.

Vinay Kumar Mittal and B. Yegnanarayana [8] conducted a study focusing on shouted speech and its production mechanisms, specifically examining changes in vocal fold vibrations and their impact on acoustic properties. The research involved 17 speakers across various loudness levels, revealing that increased loudness extends the closed phase of the glottal cycle, influencing spectral energy distribution. Proposed features aim to detect shouted speech within continuous speech, providing insights into understanding emotions like anger.While the study successfully identified features in the speech signal that discriminate shouted from normal speech across diverse contexts and speakers, challenges remain in efficiently extracting these features for practical applications due to computational complexities. The authors analyzed shouted speech characteristics by scrutinizing glottal excitation changes at different loudness levels—soft, normal, loud, and shout. Electroglottograph signals helped identify distinct glottal vibrations in shouted speech, and innovative methods addressed limitations in inverse filtering.Key findings emphasized the closed phase behavior and the low-to-high-frequency energy ratio as crucial factors in distinguishing loudness levels. The research contributes valuable insights into the acoustic properties of shouted speech, shedding light on its distinct features and challenges in practical applications.

M.S. Rudrmurthy, V. Kamakshi Prasad, and R. Kumarswamy [9] present a novel voice activity detection (VAD) method utilizing zero-frequency filtering (ZFF) to capture speech characteristics without relying on a specific mathematical model. The approach, applicable to diverse audio signals, including animal vocalizations, involves signal processing, running mean calculation, and spectral entropy determination to identify voiced regions. This technique proves robust to noise, adaptable to other VAD algorithms, and the authors plan to explore raw waveform neural network-based modeling for supervised VAD in future work. The study, partially funded by the Swiss National Science Foundation, showcases the effectiveness and potential applications of ZFF-based VAD.

Jongseo Sohn and Wonyoug Sung [10] present a Voice Activity Detector (VAD) employing soft decision-based noise spectrum adaptation for improved noise suppression in mobile telephone systems. The proposed generalized likelihood ratio test (LRT) for voice activity detection is compared with existing methods, demonstrating the reliability of the average sub-band signal-to-noise ratio (SNR) as an approximation for the Itakura-Saito distortion measure in VAD. The study also addresses noise statistic estimation to mitigate the impact of speech signals on the noise estimate.

D. Govind, S. R. M. Prasanna, and Debadatta Pati [11] present an improved method for epoch extraction in high-pass filtered speech. Focusing on Zero Frequency Filtering (ZFF) of the Hilbert Envelope (HE), the modified ZFF algorithm accommodates the characteristics of high-pass filtered speech to enhance the accuracy of epoch detection. The paper demonstrates that ZFF of the HE can more accurately identify epochs in high-pass filtered speech compared to traditional methods, particularly noted in the evaluation on telephone speech.

Ravishankar Prasad and Ekalavya Sarkar [12] propose a novel Voice Activity Detection (VAD) method utilizing empirical mode decomposition (EMD) to identify and extract the characteristic intrinsic mode function (CIMF) from speech data. The method leverages a zero-frequency filter-assisted peaking resonator (ZFFPR) to detect CIMF, improving VAD, especially in noisy conditions. Tested on clean and noisy speech data, the approach shows encouraging results in identifying voiced regions even in the presence of significant noise. This innovative method overcomes mode mixing issues in EMD and proves effective in noisy environments.

## III. CONCLUSION *(HEADING 5)*

The project aims to address the persistent challenge of accurate voice activity detection in noisy environments. By combining innovative methods and technologies, the project has achieved improved VAD performance in transitions between noise and speech, enhanced VAD in noisy environments, and robustness to noise. The developed technique has the potential to set new standards in speech processing and empower various applications, including voice recognition systems and telecommunications, with a reliable approach to voice activity detection.

## REFERENCES

[1] M. Tejus Adiga; Rekha Bhandarkar, "Improving Single Frequency Filtering Based Voice Activity Detection (VAD) Using Spectral Subtraction Based Noise Cancellation," SCOPES 2016, pp. 18-23, doi: 10.1109/SCOPES.2016.7955823.

[2] H.M. Chang, "Technical Challenge to VAD-Like Application in Mixed Landline and Mobile Environments," IVTTA '96 Proceedings, pp. 77-80, doi: 10.1109/IVTTA.1996.552764.

[3] Kenji Nakayama; Shoya Higashi; Akihiro Hirano, "A Noise Spectral Estimation Method Based on VAD and Recursive Averaging Using New Adaptive Parameter for Non-Stationary Noise Environments," 2008 International Symposium on Intelligent Signal Processing and Communications Systems, Bangkok, Thailand, 2009, pp. 1-4, doi: 10.1109/ISPACS.2009.4806668.

[4] Rui Wang and Nan Li, "Robust Voice Activity Detection Using a Masked Auditory Encoder Based Convolutional Neural Network," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 6828-6832, doi: 10.1109/ICASSP39728.2021.9415045.

[5] Wenpeng Mu and Binshan Liu, "Voice Activity Detection Optimized by Adaptive Attention Span Transformer," In IEEE Access, vol. 11, pp. 31238-31243, 2023, doi: 10.1109/ACCESS.2023.3262518.

[6] Mohammad Shahid and Cigdem Beyan, "S-VVAD: Visual Voice Activity Detection by Motion Segmentation," 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2021, pp. 2331-2340, doi: 10.1109/WACV48630.2021.00238.

[7] D. Pravena and Govinda D, "Expressive Speech Analysis for Epoch Extraction Using Zero Frequency Filtering Approach," 2016 IEEE Students' Technology Symposium (TechSym), Kharagpur, India, 2016, pp. 240-244, doi: 10.1109/TechSym.2016.7872689.

[8] Vinay Kumar Mittal and B. Yegnanarayana, "Effect of Glottal Dynamics in the Production of Shouted Speech," The Journal of the Acoustical Society of America 133 5 (2013): 3050-61.

[9] M.S. Rudrmurthy, V. Kamakshi Prasad and R. Kumarswamy, "Voice Activity Detection Algorithm Using Zero Frequency Filter Assisted Peaking Resonator and Empirical Mode Decomposition," Journal of Intelligent Systems, vol. 22, no. 3, 2013, pp. 269-282. https://doi.org/10.1515/jisys-2013-0036.

[10] Jongseo Sohn and Wonyoug Sung, "Voice Activity Detector Employing Soft Decision Based Noise Spectrum Adaptation," Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181), Seattle, WA, USA, 1998, pp. 365-368 vol.1, doi: 10.1109/ICASSP.1998.674443.

[11] D. Govind, S.R.M. Prasanna and Debadatta Pati, "Epoch Extraction in High Pass Filtered Speech Using Hilbert Envelope," Proc. Interspeech 2011, 1977-1980, doi: 10.21437/Interspeech.2011-520.

[12] Ravishankar Prasad and Ekalavya Sarkar, "Unsupervised Voice Activity Detection by Modeling Source and System Information Using Zero Frequency Filtering," Proc. Interspeech 2022, 4626-4630, doi: 10.21437/Interspeech.2022-10535.

.