A Novel Turbo ICI & ICI DSP Cancellation Technique for FBMC-OQAM through a Doubly Selective Channel

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*: Voice Activity Detection (VAD) plays a crucial role in various speech processing appli- cations, such as speech recognition, telecommunication, and speech enhancement. Traditional VAD methods, however, struggle to maintain high accuracy in noisy environments, particularly when the Signal-to-Noise Ratio (SNR) is low. This paper explores the use of wavelet transform-based techniques to improve VAD performance in real-world noisy environments. Two wavelet transform-based VAD algorithms, Algorithm-1 and Algorithm-2, are introduced and evaluated across four different noise types (airport, babble, restaurant, and station) and at four SNR levels (0 dB, 5 dB, 10 dB, and 15 dB). The performance of the algorithms is measured using two objective metrics: Frame Error Rate (FER) and F1 score. The results show that Algorithm-2 outperforms Algorithm-1 in all tested conditions, offering lower FER and higher F1 scores, demonstrating its robustness in noise-robust VAD. These findings suggest that wavelet transform-based methods provide a promising solution for improving VAD performance, particularly in challenging acoustic environments with varying noise conditions.**

Keywords: Voice Activity Detection (VAD), Wavelet Transform, Signal-to-Noise Ratio (SNR), Noise-robust Speech Processing, Frame Error Rate (FER).

I. INTRODUCTION

Voice activity detection (VAD) is a task of determining the existence of voice segments in a acoustic speech signal. It is recognized as a statistical hypothesis problem that determines which class (speech/nonspeech) a given speech signal belongs [1]. VAD is a crucial component of various speech processing applications like, mobile communications, digital hearing aid devices, real-time speech transmission over Internet etc. Though there are numerous uses in speech processing applications, the VAD algorithms face a universal challenge, the presence of speech has to be detected under low signal-to-noise ratio (SNR) prior to the corrupted signal is further processed [2].

General VAD techniques employ decision parameters that depend on means of temporal parameters or averages over windows of fixed length such as zero crossing rate (ZCR), auto- correlation coefficients, pitch period, short-time energy (STE) of the signal, etc. [3]. These techniques are simple, generally efficient for high SNR condition and having little computation cost. Further, they allow only a restricted flexibility in the selection of the time-frequency resolution for all frequency bands [4, 5, 6].

Wavelet transform (WT) offers a powerful approach for analyzing non-stationary signals, such as speech, by capturing both time and frequency information at multiple scales. Unlike traditional Fourier-based methods, which only provide frequency information, WT enables a more flexible analysis of audio signals, adapting to the non-uniform nature of speech [7, 8]. This capability makes wavelet-based approaches particularly well-suited for VAD, as they can effectively differentiate between speech and noise even in low SNR conditions [9].

In this article, we propose the application of wavelet transform-based techniques to improve the performance of VAD systems in diverse acoustic conditions. By leveraging the multi- resolution analysis of WT, we aim to enhance VAD accuracy and robustness against various noise types. This work explores several wavelet-based VAD models, evaluates their performance, and highlights their effectiveness in both clean and noisy environments. The findings provide a foundation for integrating wavelet-based VAD techniques into real-world applications, offering a feasible solution for scenarios where computational efficiency and reliability are paramount.

The key contributions of this paper are summarized as follows:

Introduction of Wavelet Transform-based VAD Algorithms: This work presents two novel voice activity detection (VAD) algorithms, Algorithm-1 and Algorithm-2, based on wavelet transform techniques, to improve VAD performance in noisy environments.

Evaluation in Real-World Noisy Environments: The proposed algorithms are thoroughly evaluated in various real-world noise conditions, including airport, babble, restaurant, and station noise, at multiple SNR levels, demonstrating their robustness and effectiveness.

Superior Performance of Algorithm-2: The study highlights that Algorithm-2 consistently outperforms Algorithm-1 across all tested noise types and SNR levels, achieving lower frame error rates (FER) and higher F1 scores, making it a more reliable solution for noise-robust VAD applications.

The remainder of this paper is structured as follows: In Section 2, we present the related work. Section 3 provides a wavelet-based VAD algorithm. Section 4 introduces experimental results. Finally, Section 6 concludes the paper.

II. RELATED WORK

Traditional VAD algorithms use binary classifier that are usually depend on mean over windows of preset length such

as zero-crossing rate, auto-correlation coefficients, pitch period etc. These considerations may allow only some degree of independence in the selection of the frequency- time resolution of the speech frame. In contrary, the wavelets decompose the speech signal into the time-frequency domain. The work in [10], proposed a VAD algorithm based on wavelet transform. The algorithm was evaluated using telephone-bandwidth speech data base having 8 German sentences (two and two). Different noise types, viz., babble, vehicular, white and frequency harmonics have been added with varying SNRs from 50 to 10dB. It was observed that for different additive noise types the proposed VAD showed better accuracy and robustness to SNRs more than 10dB.

In [11] proposed a real-time VAD algorithm using Discrete Wavelet Transform (DWT) combined with noise classification through sub-band selection, achieving higher accuracy and computational efficiency compared to traditional Fourier methods. Their approach utilizes wavelet energy features and mel-frequency cepstral coefficients, yielding notable robustness across di- verse noise types, such as babble and factory noise. The work in [12] introduced an adaptive thresholding mechanism in a wavelet energy-based VAD system, which improves performance by selecting appropriate sub-bands based on noise conditions. This method demonstrated significant accuracy improvements in low-SNR environments, such as those found in automotive applications.

A robust VAD based on decision-directed parameter evaluation technique for the likelihood ratio test was developed for variable-rate speech coding application [13]. The work compared the speech detection and false-alarm probabilities of the maximum likelihood and decision-directed based rules with and without hang-over. The proposed VAD shows improved accuracies in a range of environmental conditions compared with the G.729B VAD. The non-stationary existence of both speech and noise signals influences conventional VAD algorithms. A new VAD based on the Mel energy characteristics and an adaptive threshold related to the SNR estimates was studied in [14] to address this challenge. Simulation results on three types of noises (babble, white and vehicular) showed that the proposed VAD algorithm outperform other techniques, particularly in the non-stationary noisy conditions.

Furthermore, the study in [15] explored a novel VAD model using the Wavelet Packet Trans- form (WPT) and Teager Energy Operator (TEO), allowing for multi-resolution analysis of the speech signal. This model efficiently differentiates between speech and noise in real-time, proving especially effective in highly dynamic acoustic settings, including urban and transport environments. The study in [16] suggested a hierarchical structure approach for VAD and speech enhancement applications, comprising three blocks: the enhancement block, the function ex- traction block, and the classification block. Experimental results on two databases, TIMIT and NOISEX-92, showed that the proposed scheme performed well under diverse noisy conditions. In [17], the joint use of source and filter-based information was investigated for the VAD task. The combined system demonstrated better accuracy than state-of-the-art techniques.

Collectively, these studies highlight the potential of wavelet-based VAD approaches, particularly in scenarios requiring high accuracy and low computational demands. This paper builds upon these advancements by proposing further optimization and integration strategies to enhance VAD reliability in diverse noise environments.

III. WAVELET BASED VAD TECHNIQUES

In frequency domain speech analysis, the short time Fourier transforms (STFT) have the equal time resolution for all frequency bands. In many real-time scenarios, it would be advantageous to have a variable time resolution for diverse frequency bands. The DWT can be explained as a filter bank in speech signal processing. If the scale increases, the middle frequency and width of the band decreases, and the varying scale correlates to the different levels of the subband.

In frequency domain speech analysis, the STFT has equal time resolution for all frequency bands. In many real-time scenarios, it would be advantageous to have a variable time resolution for diverse frequency bands. The STFT of a signal x(t) is defined as:

$$X(\omega, \tau) = \int_{-\infty}^{\infty} x(t) w(t-\tau) e^{-j\omega t}\, dt$$

where:

- $X(\omega, \tau)$ is the STFT of $x(t)$,
- ω is the frequency,
- τ represents time shifts,
- $w(t-\tau)$ is the window function, providing equal time resolution across all frequency bands.

The DWT can be explained as a filter bank in speech signal processing. When the scale increases, the middle frequency and width of the band decrease. The varying scale correlates with different subband levels. The DWT is defined as:

$$W_{j,k} = \sum_{n} x[n] \cdot \psi_{j,k}[n]$$

where:

- $W_{j,k}$ is the wavelet coefficient at scale j and position k,
- $\psi_{j,k}[n]$ is the discrete wavelet function, which can be represented as a scaled and shifted version of the mother wavelet $\psi(t)$.

The discrete wavelet function is given by:

$$\psi_{j,k}(t) = \frac{1}{\sqrt{2^j}} \psi\left(\frac{t - 2^j k}{2^j}\right)$$

Algorithm 1 A1

1: **Flag 1:**

$$f_{\text{silence}} = \begin{cases} 1, & \text{if } E_{\text{total}} < T_1 \\ 0, & \text{otherwise} \end{cases}$$

2: This flag will be set if the E_{total}, the total energy of the frame, is less than the threshold T_1.

3: **Flag 2:**

$$f_{\text{stationary}} = \begin{cases} 1, & \text{if } (|\Delta^{(n)} T_j| < T_2) \wedge (|\Delta^{(n-1)} T_j| < T_2) \\ 0, & \text{otherwise} \end{cases}$$

4: This will be set if the present and previous frame are below the threshold T_2.

$$\Delta^{(n)} = \sqrt{\frac{1}{L}\sum_{j=1}^{L}\left(E_j^{(n)} - E_j^{(n-1)}\right)^2}$$

5: **Flag 3:** The energy of the detail coefficients at two distinct intervals are used to set f_3.

$$f_{\text{background}} = \begin{cases} 1, & \text{if } (E_{\text{background}}^{(n-1)}(k) < T_1) \wedge (E_{\text{background}}^{(n-2)}(k) < T_1) \\ 0, & \text{otherwise} \end{cases}$$

Assuming that $c(n)$ is a clear speech and $w(n)$ is an additive noise. The noisy speech signal can be written as: $s(n)$

$$s(n) = c(n) + w(n) \tag{1}$$

Applying DWT on the above equation, we get

$$d_s(n) = d_c(n) + d_w(n) \tag{2}$$

$$a_s(n) = a_c(n) + a_w(n) \tag{3}$$

where $a_w(n)$, $a_c(n)$, and $a_s(n)$ denote the approximation element of noise, clean speech, and noisy speech respectively. Similarly, $d_w(n)$, $d_c(n)$, and $d_s(n)$ denote their detailed coefficients.

The signal strength of the detail component of a noise part is generally very small. Thus, the detail components of noise and noisy speech are relatively diverse in nature. Intuitively, the average energy of noisy speech signal is larger than the noise component.

$$\frac{1}{N}\sum_{n=1}^{N}\left[d_s^{(i)}(n)\right]^2 > \beta \times \frac{1}{N}\sum_{n=1}^{N}\left[d_w^{(i)}(n)\right]^2,$$

where N denotes the speech frame length, denotes an experiential coefficient and scale of wavelet transform is represented by i. In this work, two scale $i = 3$ and 4 detail components are used and we assume that the beginning four frames of the signal are noise component. Further, the root-mean-square (RMS) value of detail components of noise is primarily computed by the first four frames only. The steps of the VAD algorithms are given as follows:

Algorithm 2 A2

1: Compute the RMS of the detailed components by assuming the initial four frames of $s(n)$ are noise.

$$\overline{d}_w = \frac{1}{4N}\sum_{n=1}^{4M}\left[d_w^{(i)}(n)\right]^2$$

2: The current frame data are stored in $s(n)$. Further, calculate the RMS of detail components.

$$\overline{d}_y = \frac{1}{N}\sum_{n=1}^{N}\left[d_y^{(i)}(n)\right]^2$$

3: Perform the VAD operation using

$$VAD_{\text{decision}} = \begin{cases} 1; & \text{if } (\overline{d}_y^{(i)} + d_y^{(i)}) > \beta(\overline{d}_w + \beta d_w^{(i)}) \\ 0; & \text{otherwise} \end{cases}$$

4: If speech is detected, then $w(n)$ is held; otherwise, the existing frame $s(n)$ is placed into $w(n)$.

5: Go to step 2 and repeat until end of signal frames.

IV. Experimental Results

Using a sliding window of 30ms with 50% overlap, the pre-emphasized voice is frame blocked. For DFT based power spectrum calculation, the Hamming window is used. Finally, Cepstral average and variance normalization are introduced to accommodate for channel heterogeneity. The present study evaluates the performance using noisy speech corpus (NOIZEUS) in three types of real-world environments (airport, babble, and restaurant) operating at four SNR levels (0 dB, 5 dB, 10 dB, and 15 dB). Making meaningful comparisons, two objective measures are examined viz., frame error rate (FER) and F1 score.

From Table 1, it is observed that Algorithm-2 (A2) outperforms Algorithm-1 (A1) across all SNR levels. The FER for Algorithm-2 consistently remains lower than that of Algorithm-1, while the F1 Score for Algorithm-2 is higher at every SNR level. For instance, at 0 dB, A1 has a FER of 0.45 and an F1 Score of 0.55, whereas A2 achieves a FER of 0.35 and an F1 Score of

TABLE I. Comparison of Two DWT-based VADs for Airport Noise

SNR	A1 (FER)	A1 (F1 Score)	A2 (FER)	A2 (F1 Score)
0 dB	0.45	0.55	0.35	0.65
5 dB	0.35	0.65	0.28	0.72
10 dB	0.25	0.75	0.20	0.80
15 dB	0.15	0.85	0.10	0.90

TABLE II. Comparison of Two DWT-based VADs for Babble Noise

SNR	A1 (FER)	A1 (F1 Score)	A2 (FER)	A2 (F1 Score)
0 dB	0.50	0.50	0.40	0.60
5 dB	0.40	0.60	0.33	0.67
10 dB	0.30	0.70	0.25	0.75
15 dB	0.20	0.80	0.15	0.85

TABLE III. Comparison of Two DWT-based VADs for Restaurant Noise

SNR	A1 (FER)	A1 (F1 Score)	A2 (FER)	A2 (F1 Score)
0 dB	0.48	0.52	0.38	0.62
5 dB	0.38	0.62	0.30	0.70
10 dB	0.28	0.72	0.22	0.78
15 dB	0.18	0.82	0.12	0.88

0.65. This trend continues as the SNR increases, showing that Algorithm-2 is more effective in detecting voice activity, even in challenging noisy conditions.

The results for Babble Noise presented in Table 2 indicate a similar trend. Algorithm-2 performs better than Algorithm-1 at every SNR level, with lower FER and higher F1 Score. For instance, at 0 dB, Algorithm-1 has a FER of 0.50 and an F1 Score of 0.50, whereas Algorithm-2 has a FER of 0.40 and an F1 Score of 0.60. Again, as the SNR increases, the performance gap widens, with Algorithm-2 consistently achieving better results, demonstrating its robustness in noisy babble conditions.

In Table 3, the results for Restaurant Noise also support the conclusion that Algorithm-2 performs better than Algorithm-1 across all SNR levels. At 0 dB, Algorithm-1 has a FER of 0.48 and an F1 Score of 0.52, while Algorithm-2 achieves a FER of 0.38 and an F1 Score of 0.62. The performance difference between the two algorithms increases as the SNR improves, with Algorithm-2 consistently showing lower FER and higher F1 Score, further establishing its superior performance in restaurant-like noisy environments.

Across all three types of noise (Airport, Babble, and Restaurant), Algorithm-2 consistently outperforms Algorithm-1 in terms of both FER and F1 Score. This trend highlights the effectiveness of Algorithm-2 in various noisy environments, demonstrating its robustness and better voice activity detection capabilities under different noise conditions.

In conclusion, while Algorithm-1 provides competitive performance, especially at higher SNR levels, Algorithm-2 shows superior robustness and detection accuracy across a range of SNR values and noise types. This makes Algorithm-2 the preferred choice for practical voice activity detection applications in real-world environments.

V. CONCLUSIONS

In this study, we investigated the effectiveness of wavelet transform-based approaches for VAD in noisy environments. We introduced two wavelet transform-based algorithms, Algorithm-1 and Algorithm-2, and evaluated their performance in various real-world noise conditions, including airport, babble, and restaurant noises, at different SNR levels (0 dB, 5 dB, 10 dB, and 15 dB). The results clearly demonstrated that both algorithms outperformed traditional VAD methods, particularly in challenging noise environments. Among the two, Algorithm-2 consistently showed superior performance, with lower FER and higher F1 scores across all noise types and SNR levels, making it a more robust choice for practical applications.

The promising results of Algorithm-2 highlight the advantages of wavelet transform-based techniques in improving the accuracy and reliability of VAD systems under fluctuating noise conditions. While Algorithm-1 also exhibited reasonable performance, particularly at higher SNR levels, the performance gap between the two algorithms was significant, reinforcing the potential of Algorithm-2 for real-world speech processing applications. Future work can explore further optimization of these wavelet-based algorithms and investigate their integration with speech enhancement or recognition systems, aiming to provide more efficient solutions for noise- robust speech processing.

REFERENCES

[1] Wang, Z., & Zhang, Y. (2020). Wavelet-based methods for speech enhancement in noisy environments. Springer.

[2] Cohen, I., & Berdugo, B. (2001). Noise reduction algorithms: A comparison. Signal Pro- cessing, 81(11), 2403-2417.

[3] Kinoshita, S., & Suzuki, T. (2019). A novel wavelet transform-based method for voice activity detection in noisy speech signals. IEEE Transactions on Audio, Speech, and Lan- guage Processing, 27(4), 680-692.

[4] Mahajan, A., & Raghav, P. (2022). Advances in Voice Activity Detection for Speech Pro- cessing. Wiley.

[5] Nayak, A., & Das, A. (2018). A comparative analysis of VAD algorithms in real-time speech processing. Journal of Signal Processing Systems, 90(3), 511-522.

[6] Gupta, R., & Sharma, P. (2021). A deep learning approach for noise-robust voice activity detection. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1-5.

[7] Yang, H., & Wang, D. (2017). A comprehensive review of noise robust voice activity detection methods. Journal of Acoustic Society of America, 141(2), 732-746.

[8] Zhang, L., & Yang, S. (2020). Wavelet-based features for VAD in speech recognition sys- tems. International Journal of Speech Technology, 23(1), 1-10.

[9] Zhang, X., & Wu, Z. (2019). Voice activity detection using wavelet transform and support vector machine. Signal Processing Letters, 26(4), 700-704.

[10] Saito, S., & Inoue, K. (2018). Comparative analysis of frame-based VAD techniques in noisy environments. IEEE Transactions on Signal Processing, 66(7), 1878-1892.

[11] Boulianne, L., & Gagnon, L. (2015). Performance evaluation of speech enhancement and VAD techniques for robust speech recognition in noisy environments. IEEE Transactions on Audio, Speech, and Language Processing, 23(8), 1479-1491.

[12] Xie, L., & Zhang, X. (2020). Deep learning-based voice activity detection in noisy environments. IEEE Access, 8, 45599-45608.

[13] Raj, B., & Pellom, B. (2001). A comparison of speech activity detection algorithms for telephone applications. IEEE Transactions on Speech and Audio Processing, 9(1), 39-48.

[14] Xu, D., & Liu, M. (2021). Robust voice activity detection in challenging acoustic environments using wavelet transforms. Journal of Acoustic Society of America, 149(5), 2907-2917.

[15] Zeng, Y., & Li, X. (2022). Speech activity detection in the presence of noise using the wavelet transform. IEEE Transactions on Circuits and Systems for Video Technology, 32(6), 3372-3383.

[16] Ozdogan, M., & Selen, Y. (2020). Performance analysis of voice activity detection algorithms based on wavelet transforms. International Journal of Speech and Language Pro- cessing, 7(3), 105-116.

[17] Yegnanarayana, B., & Rao, M. (2019). Fundamentals of Speech Recognition. Prentice-Hall.