Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 8, 1461-1473

Statistical Model-Based Tamil Stuttered Speech Segmentation Using Voice Activity Detection

Manjutha M[1] , Subashini P[2]

[1]*Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India*
[2]*Professor, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India*
[1]*manjutham@gmail.com*, [2]*parthasarathysubashini@gmail.com*

ABSTRACT

Speech is one of the special aid to communicate between humans. A complex speech disorder anointed stuttering is represented by repeated sounds, uttering long syllables or sentences, and blocks of speech. This instinctual speech impairment shows significant issues with the typical fluency and flow of speaking. Voice Activity Detection (VAD) is a vital front-end pre-processing method utilized in numerous speech and signal processing applications to estimate speech presence or absence in short segments of stutter speech intervals. The standard VAD is desirable to extract the stuttered speech signal features through Frame Energy, Zero Crossing Rate and autocorrelation, which detect voiced, Unvoiced or Silence (VUS) signals. Nature-inspired Particle Swarm Optimization (PSO) algorithm is proposed to detect the active Tamil speech using the optimized VAD of the different speakers of Normal Articulate Speech (NAS), Moderate Stutter Speech(MSS) and Severe Stutter Speech (SSS). This primary objective is to update the energy threshold using standard and PSOVAD methods. The proposed PSOVAD performance has been evaluated using objective benchmarks, including Front End Clipping (FEC), Mid-Speech Clipping (MSC), Over Hang (OVER), and Noise Detected as Speech (NDS). According to the experimental findings, PSOVAD can effectively isolate NAS, MSS and SSS into voiced, unvoiced, and silent under low SNR circumstances.

Keywords: Voiced, Unvoiced, Silence, Voice Activity Detection, Feature Extraction, threshold, Zero Crossing Rate, Particle Swarm Optimization

I. INTRODUCTION

Speech is a non-stationary signal, yet it is almost constant for a short period, such as 10 to 20 ms. Human communication consists of a sequence of contiguous segments of speech, non-speech and silence. The audio has speech-to-signal ratio segments, adversely impacting the systems' performance using speech-processing techniques. The effectiveness of the speech recognition system and identification of speech even in noisy environments has been detected and enhanced by using a robust Voice Activity Detection (VAD) method [2][6][9][21]. Many speech processing applications, like Voice coding, speech recognition, noise cancellation, speech enhancement, and speech synthesis, significantly utilize VAD to identify the active speech in an existing signal. Based on the excitation mode of the speech signal, it is categorized into three distinct classes: voiced, unvoiced and silent regions [26].

The human utterance of voice speech consists of remarkable stability in the spectrum. During the utterance of vowels, Constant frequencies are typically produced. The voiced speech region is generated when vibrating glottis resonates through the vocal tract and creates periodic air pulses. Depending on the structure of the vocal tract, periodic pulses are produced at specific frequencies, and the prominent part of the speech is voiced. The intelligibility of speech adequately requires Voiced speech. The passage of air pronounces the consonants is hindered by the slight constriction of the vocal tract, resulting in non-periodic sounds. Therefore, voiced speech signals have been recognized and isolated due to their regularity. A constriction is formed along the vocal tract, and the air is forced

through it at a high enough velocity to create turbulence and reproduce fricative or unvoiced sounds. Plosive sounds are produced when a complete closure is made and rapidly released.

Voice Activity Detection plays an indispensable role in acoustic detection and segmentation of the speech signal into various types, such as voice, unvoiced and silence. The basic strategy used by most VADs nowadays consists of the decision part from the feature extraction phase. The active speech and non-speech have been identified through the acoustic characteristics extracted from the feature extraction method. The various classical acoustic parameters exist in the speech processing application, such as spectral difference, short-time energy, pitch period, zero-crossing rates and normalization [22]. All the extracted acoustic features are applied to make the specific decision that results in VAD segmentation. The decision rules could be a convoluted statistical model or uncomplicated threshold values. Fig.1. represents the block diagram of the conventional VAD method.

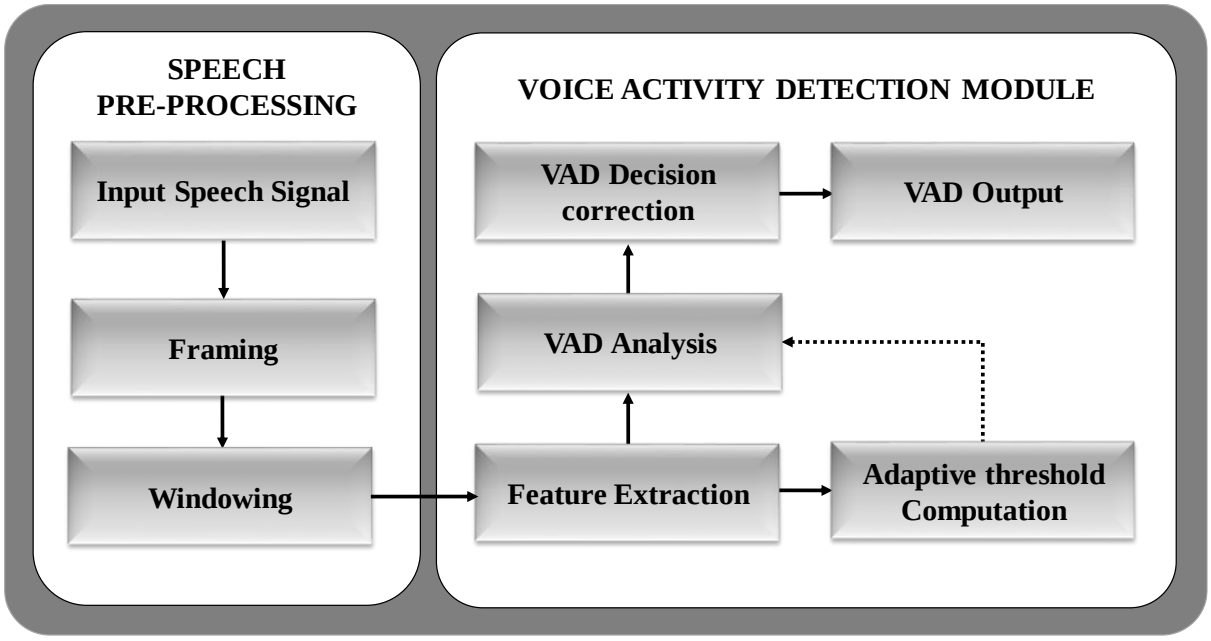


Fig.1. Block diagram of Conventional Voice Activity Detection method

The rest of the proposed work is structured as follows, Section III presents the corresponding work done by several authors linked to Voice Activity Detection. The proposed PSO-based Voice Activity Detection approach for discriminating between voiced and unvoiced stuttered speech is given in section IV. The experimental results and performance metrics are presented in section V. Finally, Section VI, concludes the paper with a summary and discussions, including future enhancement.

II. RELATED WORK

This section briefly reviews the related work performed in Voice Activity Detection in This section briefly reviews the work performed in the broad area of the speech recognition system employing Voice Activity Detection and related to the performance evaluation of various statistical measures.

Ephraim and Malah (1984) focus on a statistical model-based VAD method which increases detection accuracy and computational complexity in speech enhancement applications [27].

A Gaussian statistical model and a decision rule-based Voice Activity Detection method have been introduced using the geometric mean of the likelihood ratio (LR), which can be termed a heuristic method [11]. The originality of statistical model-based VAD has been thoroughly explored, and numerous studies have offered improved ways based on

the LR test according to two assumptions for speech presence and absence [2].

Shen, G., & Chung, H. Y. (2010) reveals the likelihood ratio based on a statistical model using Unvoiced feature normalization (UFN) that distinguishes the speech signal presence and absence. The various feature parameters like Zero Crossing Rate, Spectral Energy and Linear Prediction Coefficient were extracted to detect active speech [24]. Because the silent region energy level with a low signal-to-noise ratio (SNR) affects the classification accuracy for both voiced and unvoiced speech signals, the typical UFN technique decreases recognition performance [3]. As a result, this work explains a vital speech characteristic segmentation and identification strategy for noisy environments using UFN feature extraction techniques that classify speech signals into voiced and unvoiced. To overcome the challenges in the existing method Particle Swarm Optimization is introduced to optimize the threshold, which automatically enhances the performance of Voice Activity Detection by filtering out false detections and rejections.

III. METHODOLOGY

The proposed Voice Activity Detection method identifies the active speech and the absence of speech segments in a recorded audio signal. In order to extract the relevant features, the stuttered signals are initially pre-processed. Extracted features were input to the VAD method to take decisions using adaptive and optimized thresholds. The obtained results segment and categorizes the stuttered speech into Voiced, unvoiced and silent speech.

1) Data Acquisition

Most people living in the Tamil Nadu state of Indian country commonly speak the traditional Dravidian language, typically known as Tamil. The contemporary Tamil script has 12 vowels, 18 consonants, and one unique letter (aaytha ezutthu). A total of there are 247 characters (12 + 18 + 1 + (12 x 18)) produced when the vowels and consonants converge together, resulting in 216 compound characters. The spontaneous Tamil speech sentence has been manually formed with the help of a speech pathologist. Stutterers and fluent speakers both participated in the speech recording process. The speech was classified as moderate or severe based on the stuttering of the speakers. Six males and two females of various age groups were initially allowed to record the real-time data in .wav format using Audacity 2.1.3.

2) Speech signal pre-processing

The recorded speech signal is labelled and sampled with the frequency of 16000 kHz on the mono channel, and the signal is down-sampled to 16Khz for the subsequent analysis. The majority of methods assume that speech features are statistically steady over milliseconds. Consequently, the voiced speech signal is separated into short fixed-length using the hamming window technique, which is then treated as frames of observation to extract essential features.

3) Feature Extraction

Most features in the standard VAD technique are derived from the short-time energy, Zero-Crossing Rate (ZCR), and autocorrelation. In voice activity detection, zero-crossing rate and short-time energy are significant features to segment the stuttered speech signal region into voiced, unvoiced, and silence.

A. Short Time Energy

Non-stationary speech signals are produced from the vocal tract by time-varying amplitude and energy stimulation [13]. Therefore, it is necessary for speech processing to identify variations in energy over time that integrate with short-time speech signal areas. Voiced, unvoiced, and silent speech are classified according to their use of short-time energy because the energy associated with the unvoiced region is lower than that of the voiced region, and silence speech consists of negligible or least energy. The short-time energy of a stuttering speech signal can be computed from the overall energy using the following equation (1).

$$E_t = \sum_{x}^{\infty} S^2(x) \qquad (1)$$

where E_t is the total energy and $S(x)$ is the discrete-time signal.

The frame samples from $x = 0$ to $x = y - 1$ include the whole energy, and the computed energy of speech will be zero outside of the frame length denoted by y represented in the equation (2).

$$E_t = \sum_{x=0}^{y-1} S^2(x) \quad (2)$$

Framing and windowing after y^{th} speech frame becomes

$$S_w(x) = S(x).W(y - x) \quad (3)$$

where y is the number of samples rate or shifts at which the short-time energy is calculated for each sample and $W(y)$ is hamming windowing techniques used to estimate the time-domain parameters. Equation (4) can be used to describe the Short-Time Energy of Tamil stuttered speech signal.

$$E(x) = \sum_{x=-\infty}^{\infty} (S(x).W(y - x))^2 \quad (4)$$

B. Zero-Crossing Rate

A zero-crossing rate occurs in the discrete-time signals when successive samples have distinct algebraic signs. The frequency range of stuttered and the ordinary speech signal is computed by the rate at which zero-crossings arise [18] [23] [25]. The zero-crossing rate is the total number of times the frequency of active speech signals passes a value of zero in a specific interval of time or frame. Due to the wideband nature of speech signals, the average zero-crossing rate interpretation is substantially less accurate. Zero-crossings rate is defined as in equation (5).

$$Z(y) = \sum_{x-\infty}^{\infty} |sgn[s(x)] - sgn[s(x - 1)]| \quad (5)$$

Where, $sgn[s(x)] = \begin{cases} -1, & s(x) \geq 0 \\ 1, & s(x) < 0 \end{cases}$

The number of times a speech signal modifies the weighted average's sign within a time window is known as the zero-crossing rate. The non-stationary stuttered speech signal Zero-Crossing Rate is given by,

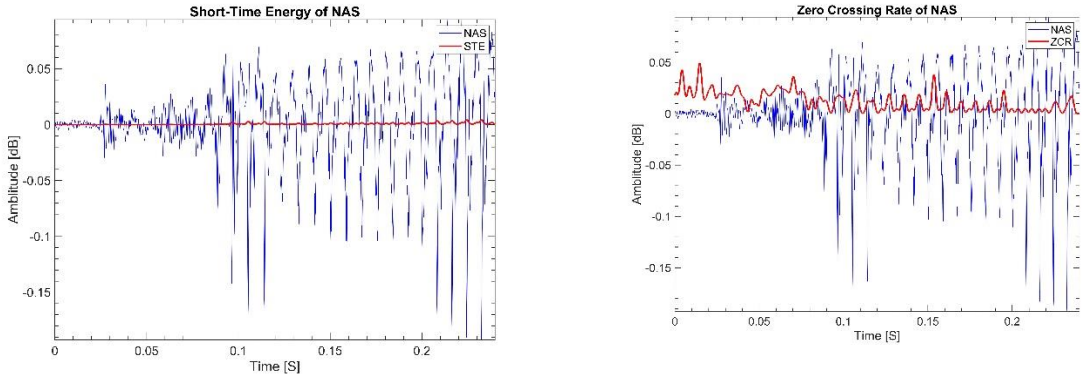
$$Z(y) = \frac{1}{2N} \sum_{x=0}^{y-1} S(x).W(y - x) \quad (6)$$

$$W(y) = \frac{1}{2N} \; if \; 0 \leq n \leq N - 1 \; or \; else = 0$$

If consecutive samples have various algebraic signs of $S(x)$ and $S(x-1)$, then the zero-crossing rate is counted and the equation (4) $|sgn[s(x)] - sgn\,[s(x-1)]| = 1$.

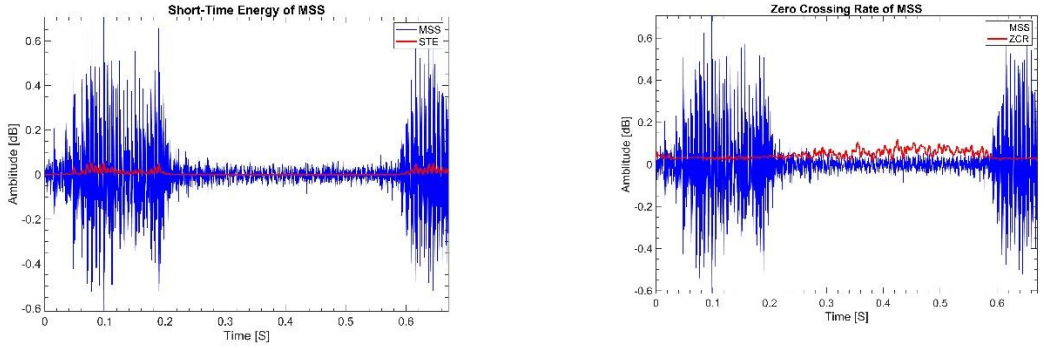
If consecutive samples have the same algebraic signs of $S(x)$ and $S(x-1)$, then the zero-crossing rate is not counted and the equation (4) $|sgn[s(x)] - sgn\,[s(x-1)]| = 0$.

The resultant short-time energy and zero-crossing rate of different speakers namely Normal Articulate Speech (NAS), Moderate Stutter Speech (MSS) and Severe Stutter Speech (SSS) articulate the Tamil word "அமுதா" (amutha) represented in Fig.2, Fig.3 and Fig.4 respectively.



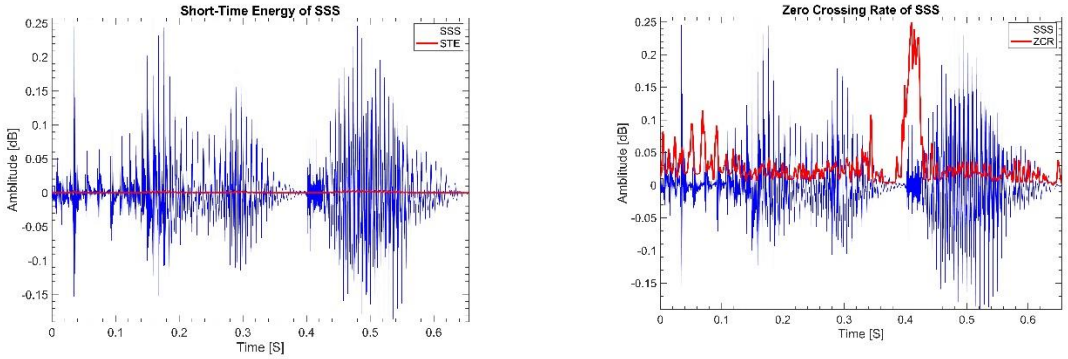
(a) **(b)**

Fig.2. (a) Short-time energy and (b) Zero crossing rate of Normal Articulate Speech (NAS) uttered the Tamil word "அமுதா" (amutha)



(a) **(b)**

Fig.3. (a) Short-time energy and (b) Zero crossing rate of Moderate Stutter Speech (MSS) uttered the Tamil word "அமுதா" (amutha)



(c) **(f)**

Fig.4. (a) Short-time energy and (b) Zero crossing rate of Severe Stutter Speech (SSS) uttered the Tamil word "அமுதா" (amutha)

C. The Normalised Autocorrelation Coefficient

The autocorrelation function is used to find the similarities between speech features with time. The correlation between adjacent samples of the signal is provided by the normalized autocorrelation coefficient function C(τ), which typically ranges between -1 and +1. Due to the frequency concentration in the low frequencies, this C(τ) value for the voiced signal is close to unity and highly correlated, but it is close to zero for the unvoiced signal.

The equation defines the normalised correlation coefficient at the unit delay (7).

$$C(\tau) = \frac{\sum_{x=\tau}^{y} S(x-\tau)}{\sqrt{\left(\sum_{x=\tau}^{y} S^2(x).\left(\sum_{x=0}^{y} S^2(x)\right)\right)}} \quad (7)$$

where $\tau = 1$

A. Auto-Correlation Periodicity

A periodic function of autocorrelation repeats indefinitely and has its periodicities. A periodic autocorrelation with undulations that decrease amplitude across lags results from a periodic signal degraded by the noise [14]. A completely uncorrelated signal will have its autocorrelation delta function centre at zero lag. The energy was zero for the uncorrelated signal and high for the periodic signal in all non-zero delays. The original signal power spectrum can be found in the Fourier transform of the autocorrelation function. The lower frequency half band contains more energy in a voiced segment's power spectrum than the higher frequency half band. A noise frame or an unvoiced frame leans toward the opposition of the voiced frame.

IV. PROPOSED METHOD

Zero crossing and energy are necessary features in Voice Activity Detection to split the speech region into voiced, unvoiced, and silent. The most significant temporal features, such as the speech energy and zero-crossing rate, have been extracted in which the threshold value has been optimized using Particle Swarm Optimization [15]. The Particle Swarm Optimization is proposed to improve active speech signal detection efficiently. Each incoming frame is used to compute the short-time energy threshold. The search space retains several particles with random initialization of location and velocity for each speech frame energy threshold, considered particles in the PSO [16]. The objective function has been computed to evaluate each particle. The particle movement is determined through the search space by integrating its previous current and best positions with one or more swarm members, [4][8] [19]. The swarm will eventually move close to the fitness function's optimal location, similar to a flock of birds foraging together. The three-dimensional search space vectors are used to alter each particle which exists [1]. The particle's best position is Pbest, whereas the best position of its neighbourhood is Gbest. The particle's current location and velocity are PP_i and PV_i, respectively. A velocity PV_i is added to the particle's current position until the condition is met that updates the particle's position each time.

The short-time energy acquires input from the neighbourhood's optimal threshold, as the PSO defines. The neighbourhood's best threshold acquired from the PSO is given as input to the short-time energy. The incoming frame energy threshold is compared to the PSO threshold energy. The speech frame is categorized as voiced speech if the produced frame energy exceeds the PSO threshold. Otherwise, it is silent or voiceless. This threshold validation is carried out to the end of the frame. After computing each frame's energy threshold, VAD eventually makes a decision. The PSO provides better prediction results compared to the traditional method. The following significant steps describe detecting the PSO-based frame energy threshold.

Step 1: Initialization involves generating particles of size m with random locations PP_i and velocities PV_i and defining parameters like the constriction coefficient and random numbers.

Step 2: The fitness function determines the best threshold among each particle, the mathematical expression of sphere objective function is represented in equation (8).

$$f(x) = \sum_{i=1}^{d} x_i^2 \quad (8)$$

Step 3: The individual best particle (Pbest) is identified from each iteration based on the fitness function (maxima) evaluation.

Step 4: Determine the ideal neighbourhood: The current particle PP_i with the best Pbest value is analyzed. If the current value is more significant than Pbest, it is assigned to the current position, and PP_i equals the current velocity PV_i. The best particle in the neighbourhood is found, and its index is assigned to the variable Gbest.

Step 5: Updating the particle's velocity and location: After each iteration, the particle's

velocity and position are updated using the mathematical equations 9 and 10.

$$PV(i+1) = \chi\,[PV(i) + \varphi_1\,(Pbest - PP(i)) + \varphi_2\,(Gbest - PP(i))] \quad (9)$$

$$PP_{(i+1)} = PP_{(i)} + PV_{(i)} \quad (10)$$

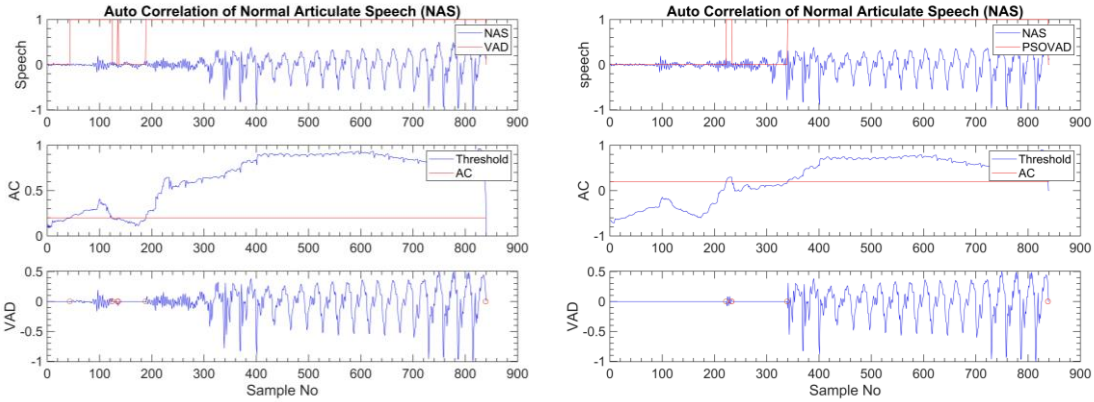
where $\chi = 2|2 - \varphi_1 - \varphi_2 - 4\varphi| = \varphi_1 + \varphi_2$,

$\varphi_1 = C_1 * R_1$ and $\varphi_2 = C_2 * R_2$

χ = constriction factor (The constriction coefficient ensures particle convergence over time while simultaneously preventing particle collapse.)
$C_1 = C_2 = 2.0$, $PV_{(i+1)}$ = velocity of i^{th} particle, $PP_{(i+1)}$ = Position of i^{th} particle, P_{best} = Personal best and G_{best} = Neighborhood (social) particle best g

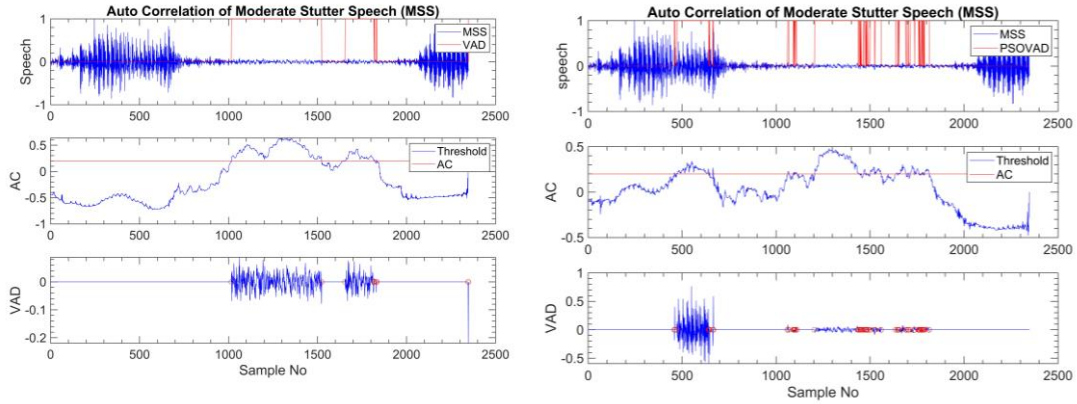
Step 6: The predetermined frame energy threshold value has been found by repeating steps 2–6 for a maximum of 1000 iterations.
Step 7: After evaluating predetermined runs, the PSO algorithm offers Pbest and Gbest potential solutions (best threshold value).
Step 8: The resulting global best threshold value is used for short-term energy processing.

The resultant autocorrelation of conventional threshold and PSOVAD threshold for Normal Articulate Speech (NAS), Moderate Stutter Speech (MSS) and Severe Stutter Speech (SSS) are represented in Fig.5., Fig.6., and Fig.7 respectively.



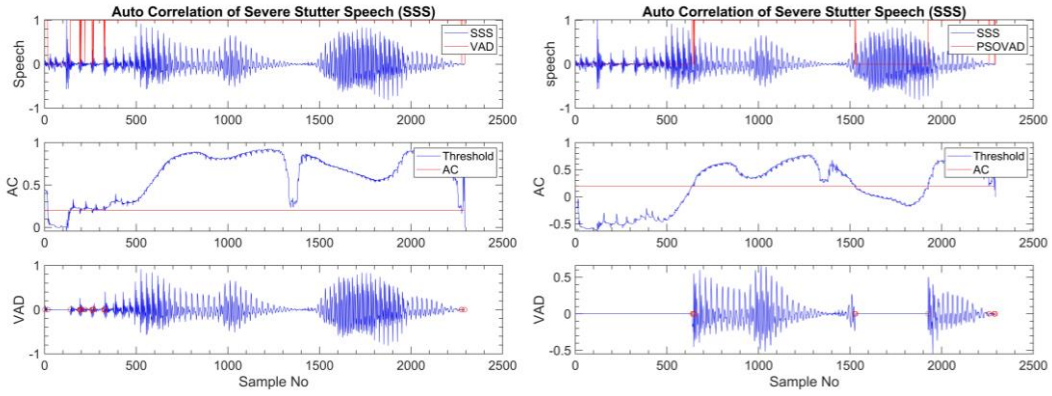
(a) **(b)**

Fig. 5. (a), (b) Auto Correlation of Normal Articulate Speech (NAS) conventional VAD and PSOVAD threshold



(a) **(b)**

Fig. 6. (a), (b) Auto Correlation of Moderate Stutter Speech (MSS) conventional VAD and PSOVAD threshold



(a) **(b)**

Fig. 7. (a), (b) Auto Correlation of Severe Stutter Speech (SSS) conventional VAD and PSOVAD threshold

1) VAD Decision

The detection module violates the VAD, and even if the detection module is resilient, a poorly built detection module will result in an unacceptable probability of error [7] [10] [17][20]. The objective parameters are used to evaluate the performance of a VAD algorithm. The optimal VAD options are obtained by manually labelling speech and non-speech times in pure voice recorded in a quiet environment. The following metrics used to compute the traditional and proposed VAD errors rate.

A . Front End Clipping (FEC)

Front End Clipping (FEC) happens when speech is misclassified as noise transitioning from the noise zone to the speech region. The following equation is used to calculate FEC (11) [5].

$$\mathrm{FEC} = \frac{\mathrm{N_F}}{\mathrm{N_{speech}}}.100 \quad (11)$$

where $\mathrm{N_F}$, is the number of samples misclassified as noise when passing from noise to speech and $\mathrm{N_{speech}}$ is the total number of samples of speech from an ideal VAD.

B. Mid-speech clipping

Mid-speech clipping occurs when speech is misclassified as noise during an utterance. The MSC measure is obtained from the equation (12).

$$\mathrm{MSC} = \frac{\mathrm{N_M}}{\mathrm{N_{speech}}}.100 \quad (12)$$

where $\mathrm{N_M}$, is the number of samples misclassified as noise during an utterance.

C. Over Hang (OVER)

OVER is the amount of noise regarded as speech during the transition from speech to non-speech or noise. The equation is used to calculate OVER (13) [12].

$$\mathrm{OVER} = \frac{\mathrm{N_O}}{\mathrm{N_{silence}}}.100 \quad (13)$$

where $\mathrm{N_O}$, is the number of samples interpreted as speech while passing from speech to silence period and $\mathrm{N_{silence}}$, is the total number of samples from the silence period of an ideal VAD.

D. Noise Detected as Speech (NDS)

NDS is a measure of noise interpreted as speech during a period of silence. The equation determines the NDS (14)

$$\mathrm{NDS} = \frac{\mathrm{N_N}}{\mathrm{N_{silence}}}.100$$

where $\mathrm{N_N}$ refers to the number of samples interpreted as speech while in the silence region.

V. RESULTS AND DISCUSSION

The proposed work is developed on the MATLAB R2013a computational platform. The objective parameter is analyzed and compared using PSO to validate the efficiency of the proposed technique. Clipping errors (FEC+MSC) are the total amount caused by front-end and mid-speech clipping. OVER and Noise Detected as Speech measurements provide the percentages of false alarms in the detected voiced and unvoiced segments, which are known as insertion errors (OVER+NDS). The experimental results of this work reveal that the temporal characteristic retrieved from the windowed data effectively improves voice activity detection.

The performance measures of Front End Clipping (FEC), Mid-Speech Clipping (MSC), Over Hang (OVER), and Noise Detected as Speech (NDS) is determined and compared with the conventional VAD method and the proposed PSO-based VAD method, which is interpreted in the following Fig.8, Fig.9, Fig.10 and Fig. 11 respectively.

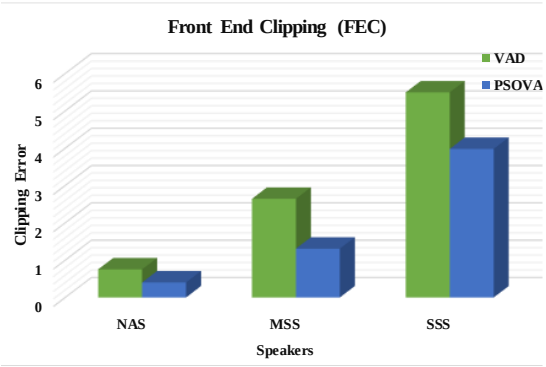


Fig. 8. Objective Parameter evaluation based on Front End Clipping (FEC)

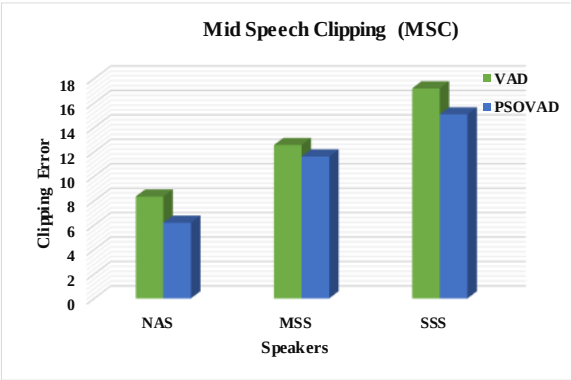


Fig. 9. Objective Parameter evaluation based on Mid Speech Clipping (MSC)

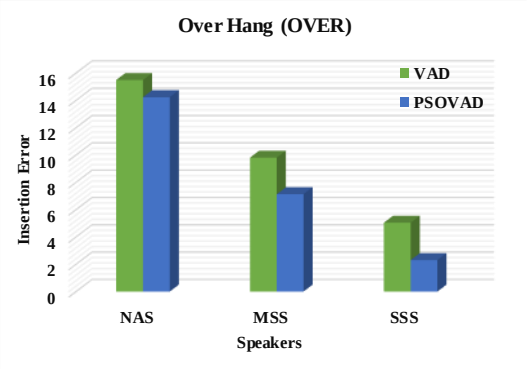


Fig. 10. Objective Parameter evaluation based on Over Hang (OVER)

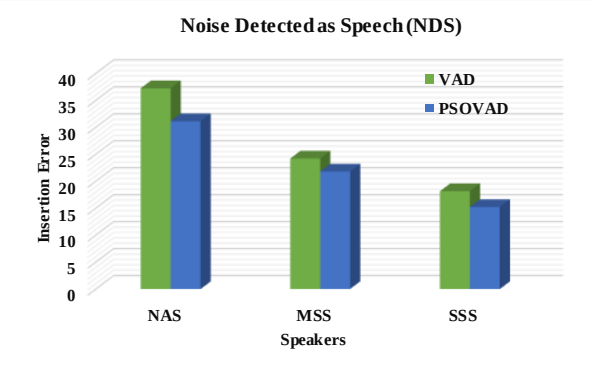


Fig. 11. Objective Parameter evaluation based on Noise Detected as Speech (NDS)

Fig. 8, Fig. 9, Fig. 10 and Fig. 11 show that the proposed PSO VAD method performs exceptionally well in maintaining robust speech intelligibility compared to conventional VAD. The FEC is a clipping occurs from the noise region to the speech region. The proposed PSOVAD method significantly reduces the FEC error 60.53%, 67.70% and 31.96% for NAS, MSS and SSS respectively. The MSC is arises due to speech misclassified as noise during the speech segmentation. From the experimental result, the proposed PSO based VAD method reduces 63.74% for NAS, 7.82% for MSS and 13.19% of overall MSC error when comparing to the standard method. During the speech processing, noise can be interpreted as speech which causes Over Hang. The proposed method decreases to 8.45% of NAS, 31.58% of MSS and 74.55% of SSS overall insertion error when comparing to the conventional method shown in Fig.10. The Noise Detected as Speech for the proposed PSOVAD method produces low insertion error for NAS 17.89%, MSS 10.47% and SSS 17.55% comparing to the classical VAD method which is shown in Fig.11. The error rate of the objective parameter was evaluated and compared with conventional and PSOVAD method. In the PSOVAD approach, the clipping errors of the FEC+MSC measure are absolutely low. Because of the few clipping errors in the proposed method, the segmentation of Voice Activity Detection outperforms well. The proposed VAD method provides good results for SNR values ranging from 5dB and higher.

VI. CONCLUSION

Voice Activity Detection is a complicated problem, and numerous authors have suggested several methods to classify active speech. A Feature Extraction, a Decision Module, and a Decision Smoother are required components of any VAD algorithm. Any VAD algorithm requires a Feature Extraction, a Decision Module, and a Decision Smoother. VAD errors can be significantly decreased by combining many features and employing optimal thresholds with a robust decision smoother. The algorithm's primary idea is to compute a set of features from the signal intended to estimate characteristics that distinguish speech from non-speech. The output of the threshold

computation determines whether the signal is speech or not. The proposed PSOVAD enhances the identification of active speech more efficiently than the traditional method for NAS, MSS and SSS.

The performance of the PSOVAD approach under noise situations is evaluated using objective parameters for both male and female speakers. As a result, the PSOVAD technique outperforms the standard VAD. In the future, the research work could be expanded to compare with other VAD methods such as linear energy-based VAD (LED), a pattern recognition approach to voiced-unvoiced categorization, and VAD based on statistical metrics, among others.

REFERENCE

[1]. Aggarwal, V., Jin, W. O., & O'Reilly, U. M. (2006). Filter approximation using explicit time and frequency domain specifications. Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation - GECCO '06, 753–760. https://doi.org/10.1145/1143997.1144132

[2]. Al-Hashemy, B., & Taha, S. (1988). Voiced-unvoiced-silence classification of speech signals based on statistical approaches. Applied Acoustics, 25(3), pp. 169–179. https://doi.org/10.1016/0003-682x(88)90092-8.

[3]. Chung, K., & Oh, S. Y. (2015). Voice activity detection using an improved unvoiced feature normalization process in Noisy Environments. Wireless Personal Communications, 89(3), 747–7591984. https://doi.org/10.1007/s11277-015-3169-5

[4]. D. Tan, W. Luo and Q. Liu, "Multi-objective particle swarm optimization algorithm for engineering constrained optimization problems," 2009 IEEE International Conference on Granular Computing, 2009, pp. 523-528, doi: 10.1109/GRC.2009.5255064.

[5]. Davis, A., Nordholm, S., & Togneri, R. (2006). Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold, IEEE Transactions on Audio, Speech and Language Processing, 14(2), 412–424. https://doi.org/10.1109/tsa.2005.855842

[6]. Eshaghi, M., & Karami Mollaei, M. (2010b). Voice activity detection based on using wavelet packet. Digital Signal Processing, 20(4), 1102–1115. https://doi.org/10.1016/j.dsp.2009.11.008

[7]. Henning Puder and Oliver Soffke, "An Approach to an Optimized Voice-Activity Detector for Noisy Speech Signals", In Proc. EUSIPCO-2002, 11th European Conference on Signal Processing, Toulouse, France, vol. 1, pp. 243-246.

[8]. J. Kennedy and R. Eberhart, "Particle swarm optimization," Proceedings of ICNN'95 - International Conference on Neural Networks, 1995, pp. 1942-1948 vol.4, doi: 10.1109/ICNN.1995.488968.

[9]. J. Tchorz and B. Kollmeier, "SNR estimation based on amplitude modulation analysis with applications to noise suppression," in IEEE Transactions on Speech and Audio Processing, vol. 11, no. 3, pp. 184-192, May 2003, doi: 10.1109/TSA.2003.811542.

[10]. Jiang, W., Lo, W. K., & Meng, H. (2010). A new voice activity detection method using maximized Sub-band SNR. 2010 International Conference on Audio, Language and Image Processing. https://doi.org/10.1109/icalip.2010.5685008

[11]. Jongseo Sohn, Nam Soo Kim, & Wonyong Sung. (1999). A statistical model-based voice activity detection. IEEE Signal Processing

Letters, 6(1), 1–3. https://doi.org/10.1109/97.736233

[12]. Korkmaz, Y., & Boyacı, A. (2022). milVAD: A bag-level MNIST modelling of voice activity detection using deep multiple instance learning. Biomedical Signal Processing and Control, 74, 103520. https://doi.org/10.1016/j.bspc.2022.103520

[13]. Liu, B., Tao, J., Mo, F., Li, Y., Wen, Z., & Liu, S. (2014). Efficient voice activity detection algorithm based on sub-band temporal envelope and sub-band long-term signal variability. The 9th International Symposium on Chinese Spoken Language Processing, pp. 531–535. https://doi.org/10.1109/iscslp.2014.6936602

[14]. M. Jalil, F. A. Butt and A. Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals," 2013 The International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), 2013, pp. 208-212, doi: 10.1109/TAEECE.2013.6557272.

[15]. M. Manjutha, Dr.P.Subashini. (2017). Particle Swarm Optimization Based Voice Activity Detection for Stuttered Tamil Speech, International Journal of Computer Engineering and Applications, 11(9), pp.1-14.

[16]. M. Manjutha, P. Subashini and M. Krishnaveni, "Analysis on Regularity of Speech Energy based on Optimal Thresholding for Tamil Stuttering Dataset," 2019 IEEE International Smart Cities Conference (ISC2), 2019, pp. 143-149, doi: 10.1109/ISC246665.2019.9071726

[17]. Ma, Y., & Nishihara, A. (2013). Efficient voice activity detection algorithm using long-term spectral flatness measure. EURASIP Journal on Audio, Speech, and Music Processing, 2013(1). https://doi.org/10.1186/1687-4722-2013-21

[18]. Mohd Hanifa, R., Isa, K., Mohamad, S., Mohd Shah, S., Soosay Nathan, S., Ramle, R., & Berahim, M. (2019). Voiced and unvoiced separation in Malay speech using zero crossing rate and energy. Indonesian Journal of Electrical Engineering and Computer Science, 16(2), 775. https://doi.org/10.11591/ijeecs.v16.i2.pp775-780

[19]. R. C. Eberhart and Y. Shi, "Comparing inertia weights and constriction factors in particle swarm optimization," Proceedings of the 2000 Congress on Evolutionary Computation. CEC00 (Cat. No.00TH8512), 2000, vol.1, pp. 84-88 doi: 10.1109/CEC.2000.870279.

[20]. Ramirez, J., M., J., & C., J. (2007). Voice Activity Detection. Fundamentals and Speech Recognition System Robustness. Robust Speech Recognition and Understanding. Michael Grimm and Kristian Kroschel(Ed.), InTech, Vienna, Austria: I-Tec., 2007, Chapter 5, pp.460. https://doi.org/10.5772/4740

[21]. Rangachari, S., & Loizou, P. C. (2006). A noise-estimation algorithm for highly non-stationary environments. Speech Communication, 48(2), 220–231. https://doi.org/10.1016/j.specom.2005.08.005

[22]. Saeedi, J., Ahadi, S. M., & Faez, K. (2013). Robust voice activity detection directed by noise classification. Signal, Image and Video Processing, 9(3), 561–572. https://doi.org/10.1007/s11760-013-0479-5

[23]. Savoji, M. (1989). A robust algorithm for accurate endpointing of speech signals. Speech Communication, 8(1), 45–60.

https://doi.org/10.1016/0167-6393(89)90067-8

[24]. Shen, G., & Chung, H. Y. (2010). Cepstral distance and log-energy based silence feature normalization for robust speech recognition. The Journal of the Acoustical Society of Korea, 29(4), 278–285.

[25]. Sunil Kumar, S., & Sreenivasa Rao, K. (2016). Voice/non-voice detection using phase of zero frequency filtered speech signal. Speech Communication, 81, 90–103. https://doi.org/10.1016/j.specom.2016.01.008

[26]. Y Padma Sai, V. K. (2015). Design and Implementation of Silent Pause Stuttered Speech Recognition System. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 04(03), 1253–1260. https://doi.org/10.15662/ijareeie.2015.0403012

[27]. Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 6, pp. 1109-1121, December 1984, doi: 10.1109/TASSP.1984.1164453