FRAMEWORK BASED SUPERVISED VOICE ACTIVITY DETECTION USING LINEAR AND NON-LINEAR FEATURES

G.B.Gour

Assistant Professor, Department of ECE, BLDEAs V.P.Dr.P.G.Halakatti College of Engineering and Technology, Vijayapur- 586101, Karnataka, INDIA
ec.gururaj@bldeacet.ac.in

Dr. V.Udayashankara

Professor, Department of Instrumentation and Technology Engineering, Sri Jayachamarajendra College of Engineering, Mysuru- 570006, Karnataka, INDIA
v_udayashankara@sjce.ac.in

Dr. Dinesh K Badakh

Professor and Head, Department of Radiation Oncology, Sri Siddhivinayak Ganapati Cancer Hospital, Miraj-416410, Maharastra, INDIA
dineshbadakh@yahoo.co.in

Dr. Yogesh A. Kulkarni

Professor, Department of Medicine, Nargis Dutt Memorial Cancer Hospital, Barshi-413401, Maharastra, INDIA
vdkulyog@gmail.com

Abstract

Voice activity detection (VAD) methods based on linear features are limited by linearity assumption and correct estimation of the pitch. The nonlinear dynamic methods can analyse irregular vocal cord behaviours and are found to be useful in the areas of voice study, clinical treatment evaluation, voice classification as per the Titze. The development of VAD using nonlinear features has required more attention. However, speech recordings in most of the practical scenarios like, in hospitals, research centers, video conferencing, and forensics are affected by babble noise. Moreover, correct estimation of signal to noise ratio (SNR) also depends on reliable voice activity detection. By looking at such challenges, the paper presents a framework based VAD using the combination of linear and nonlinear features with the two-step noise reduction (TSNR) for the possible speech enhancement. Speech segments are classified by using a supervised support vector machine (SVM). The framework is evaluated and compared with different time and frequency domain based VADs on the speech containing continuous and sustained vowels with babble noise. For this purpose, the experimental study is carried on continuous NOIZEUS corpus with babble noise at varying SNR levels. As the vocal disorder is more prominent, laryngeal pathologies based database from Saarbruecken Voice Database (SVD) and Laryngeal cancer data are used for sustained vowels. The study revealed the importance of bio-inspired linear and nonlinear features in the VAD. Finally, the proposed VAD is found to be more robust as far as vowels and continued speech are concerned.

***Keywords*: Voice activity detection; Pathological voice analysis; Vocal system; Non-linear dynamic methods; SVM.**

1. Introduction

Voice activity detection methods are proposed and studied with continuous speech corpus. The study of VAD in the case of sustained vowels will be useful as vowels play a significant role in the voice production process. Moreover, sustained phonations are independent of languages. As the recordings of pathological voices normally subjected to babble noise, the study of such sustained voice samples has to be focused on accurate identification of speech frames using appropriate VAD. SNR 's estimation is challenging as the speech itself is static, and is also highly corrupted by a variety of unknown noises. Hence, a good estimation of SNR requires reliable detection of speech segments [1]. As babble noise is having similar statistical characteristics as that of the speech, it becomes difficult to identify speech segments. Two main parts of the VAD system are discriminating features and reliable decision process. From the previous studies, the features are broadly categorized into linear and nonlinear with

subsequent time and frequency domains. The decision processes are based either on a threshold or statistical modeling and type of machine learning. The threshold-based decision method is considered useful with discriminating features, and machine learning or statistical decision-making methods are useful for higher SNR [16].

As the performance of the time domain features decreases with varying noise levels, frequency domain features were used to build VADs such as Mel-frequency cepstral coefficients (MFCC) and equivalent rectangular bandwidth (ERB) gammatone filters as in [12]. Vowel sounds having harmonic peaks with higher energies were used in building VADs showing an accuracy of 83% to 96% [11]. Formant based VAD was developed with the prior knowledge of resonant frequencies and the rate of detection was varied with the speaker and language. At lower SNR, this method was found not to be robust against non-stationary noise [9]. Threshold-independent experimental investigations were carried out with the TIMIT continuous speech corpus, QUT-NOISE, NOISEX-92 databases using short-term power and SNR features. Hence, threshold-based decisions are either empirically determined or adaptive. With noise, as the linear separability of features decreases, nonlinear methods like statistical based or machine learning-based methods were adopted. These methods were independent of the cut-off decision but necessitate training information for various types of background noises [18]. The hierarchical framework for VAD using modified wiener filtering in the speech enhancement using NOISEX-92 and TIMIT databases was proposed. The efficacy of the proposed VAD algorithm was reported between 66.8% and 97.8%, with the SNR levels rising from 0 dB to 20dB [22].

VAD having either time or frequency domain features or both are limited by its linearity assumptions. Commonly used linear methods like time, frequency, and cepstral analysis were limited by the correct estimation of the pitch. But, the vocal cords' structure and behaviour are nonlinear. The airflow down the system of the human vocal folds follows turbulent flow rules that lead to a nonlinear approach. Therefore, the application of LTI systems theory for representing the dynamics of speech production is not sufficient. The non-linear approach to speech signal processing involves, extraction of chaotic characteristics like the largest Laypunov exponent (LLE), correlation dimension (CD) which, have shown a better classification rate [10, 7, 4, 6]. More complexity and computational costs are added while developing VAD by involving high dimensional feature vector and DNN or advanced architectures of neural networks [23]. In the context of this, the paper presents the development of a reliable and compact VAD system using a combination of linear and nonlinear features. By looking at the challenges in voice activity detection and non-linearity of speech, a framework is built as explained in section 2. The section 3 depicts the methods with feature extraction, section 4 materials adopted, section 5 experimental set-up and sections 6 and 7 explains the results with conclusions.

2. Experimental study using the frame work

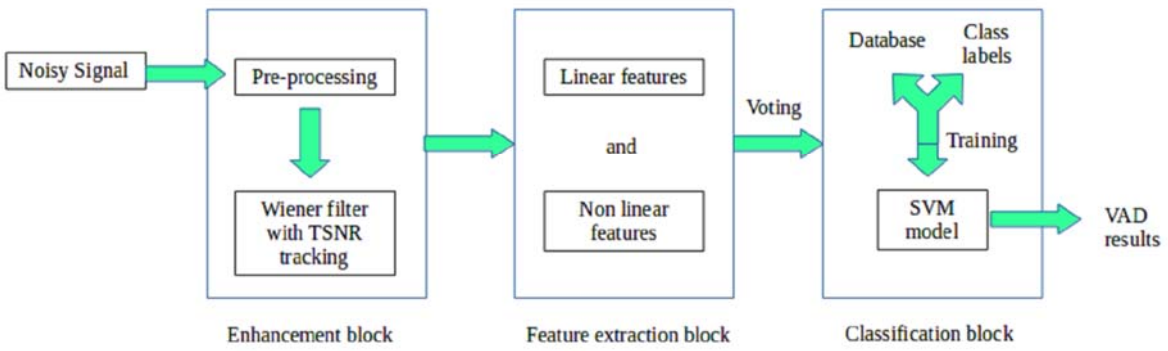
The framework is outlined with the contribution of both linear and nonlinear features as shown in Fig.1. The framework is built on a Linux platform using GNU Octave 4.0 which is similar to the licensed version of Matlab, with the following three parts.

2.1. *Pre-processing stage with Speech Enhancement*

Throughout the work, the pre-processing stage has been kept the same. At first, the speech signal is pre-emphasized using α = 0.97 using the following filter (1),

$$y_p(n)=y(n)-\alpha y(n-1) \quad (1)$$

Then, the speech enhancement is carried out using two-step noise reduction method (TSNR) as explained below.



. Fig. 1. Flow graph of the present work

2.2. *Speech Enhancement with TSNR*

The main idea of using wiener filtering is to enhance the frequencies in speech regions and to suppress the frequencies in non-speech regions [2]. In TSNR method, the disruptive speech signal, as per the additive noise model, includes the combination of voice s (t) and noise n (t), given by,

$$x(t)=s(t)+n(t) \quad (2)$$

Their k^{th} spectral component in the p^{th} frame is given by (3),

$$X(p,k)=S(p,k)+N(p,k) \quad (3)$$

As there is no direct solution available for spectral estimation of S(p,k) and is estimated by a priori and a posteriori SNR, only magnitudes are considered for SNR estimation without phase in decision directed (DD) approach. With β , the controlling factor and variance of $SNR_{prio}^{DD}(p,k)$ reduces with a variation of SNR, which is described by an over and under estimation of a priori SNR called reverberation effect. In the second step, the spectral gain G_{DD} (p, k), is used to estimate an a priori SNR at the next successive frame. This method is known as two-step noise reduction technique. Using the Hamming window, the frame length of 25 ms and frame overlaps of 50% are used throughout the work. Fig. 2, shows a significant enhancement in speech signal using the TSNR method with improvement in segmental SNR from around 0.1 dB to 0.3 dB with consequent reduction of residual noise.

2.3. *Feature extraction and Classification*

The SVM is used to classify segments between speech and non-speech regions due to low computational costs. The data comprising of n-dimensional feature vectors are first labeled by using the Audacity tool, scaled and normalized before feeding to the SVM. In all the cases, SVM having radial basis function with gamma = 0.5 and epsilon = 0.1 is used [3]. The voting is performed by the respective threshold-based methods as discussed in section 3. The 60% of data is used for training and 40% for testing purposes. The feature extraction is discussed in line with methods.

3. Methods

Owing to the objectives of the present work, an experimental investigation of the following five different VAD methods is carried out. The first four methods are purely based on linear features with different time and frequency domain features and the fifth method is purely based on nonlinear parameters as shown in Table 1.

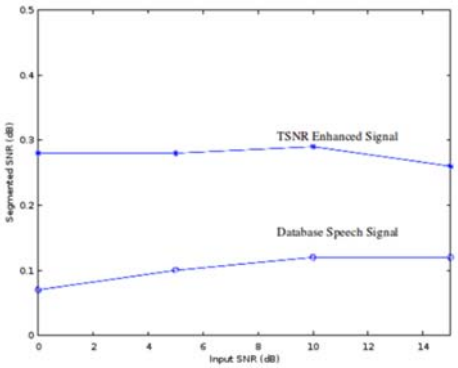


Fig. 2. Segmental SNR of noisy signal and enhanced signal using TSNR

Methods	Description	References
m1	Threshold based VAD using spectral entropy, STE, ZCR and LPC	[13,15,16,18,22]
m2	Threshold based VAD using gammatone filtering and entropy	[21]
m3	VAD using GFCC and delta, double delta features	[20,14]
m4	VAD using MFCC and delta, double delta features	[7,12]
m5	VAD using nonlinear features, mtutual information (MI), false nearest neighborhood (FNN), correlation dimension (CD), correlation entropy (CE2), Renyi entropy –order 1 and 2 (RE1, RE2), largest Laypunov exponent (LET)	[4,6,7,8,10,17]

Table 1. The notations used in the present work to identify five different methods.

Time-domain based features are used in the first two methods of VAD using decision-based on thresholding, which are explained as follows;

3.1. *Method-1 (m1)*

Since the voiced segment's short-time energy is more than the unvoiced segment, STE is easily calculated and STE is E(n) of frame *n* of the speech signal x(i) with frame length N. The zero-crossing rate (ZCR) Z_n is calculated

using sign function as in Eqn. (4). The spectral entropy shows less organization in non-speech regions as compared with speech regions. This can be computed by the Eqn. (5),

$$Z_n = \sum_{m=-\infty}^{\infty} |\text{sgn}[s(m)] - \text{sgn}[s(m-1)]| \quad (4)$$

$$H(s) = -\sum_{i=1}^{N} P\left(s(i)\right) \log_2 P\left(s(i)\right) \quad (5)$$

Where S= [s (1),..., s (I),..., s (n)] displays a *n* symbol sequence. The linear prediction error (LPE) indicates the segment of speech or non-speech. The LPE, e(n) is given mathematically by (6) between the actual sample s(n) and its projected sample;

$$e(n) = s(n) - \hat{s}(n) \quad (6)$$

The thresholding factor, including STE, ZCR, entropy and LPE is computed by (7),which is used in VAD decision.

$$D = \text{STE}(1 - \text{ZCR})(1 - \text{LPE})(\text{Spectral entropy}) \quad (7)$$

3.2. *Method-2 (m2)*

In the work of [21], an unsupervised VAD using biologically inspired gammatone filter is used, which acts like the cochlea basilar membrane vibration. The gammatone filter center frequencies, f_c, are expanded in frequency, proportional to the equivalent rectangular bandwidth (ERB) scale as shown in (8) that acts like the

$$\text{ERB} = 24.7\left(4.37X10^{-3} f_c + 1\right) \quad (8)$$

Then gammatone filterbank is constructed using 32 center frequencies by gammatone impulse response (9),

$$g(t) = \alpha \circ t^{n-1} e^{-2\pi bt} \cos(2\pi f_c t + \varphi) \quad (9)$$

which covers the useful human speech spectral bands. Each of these speech signals after passing through these gammatone filter banks, e(n) is multiplied with a weighting function w(k) to compensate for noise effects. Then, entropy H(X) is computed by using normalized and weighted signal e(k).w(k), to obtain instantaneous information, labeled as ∂(n).

$$H(x) = -\sum_{k=0}^{K} p_k \log_2 p_k \quad (10)$$

The threshold is obtained by the sum of the average and three-fold standard deviation of the lowest 20 percent of ∂(n), which gives the estimate of the high noise floor envelope.

Methods	Description	References
$\mathbf{m_6}$	VAD using nonlinear features with MFCC	[13,15,16,18,22]
$\mathbf{m_7}$	VAD using nonlinear features with MFCC and m1 features	[21]
$\mathbf{m_8}$	VAD using nonlinear features with m1 features	[20,14]

Table 2. The notations used in the present work to identify remaining three methods

3.3. *Method-3 (m3)*

Gammatone Cepstral coefficients are extracted by applying gammatone filter bank with 32 center frequencies. ERB gammatone filtering is used to suppress the randomly varying spectral components to enhance the dynamic parts of the corrupted speech. Then applying the logarithm of discrete cosine transformation (DCT) to each of the channels to get 13 gammatone Cepstral coefficients (GTCC). Total 39 dimensional vector containing GTCCand its derivatives, delta-GTCC and double delta GTCC are extracted. The supervised classification is carried out by using a SVM for speech detection [14, 20].

3.4. *Method-4 (m4)*

In this method, the 13 MFCC are derived from each of the speech frames by taking the discrete Fourier transform (DFT). In order to obtain its power spectrum, it is then passed through a triangular filter bank with 24 filters, uniformly positioned on the Mel scale. Log energy is computed for each of the banks, which is found to be

sensitive to small variations in the articulatory movements. Then the MFCC of 13 coefficients is obtained by using DCT. Total 39-dimensional vectors of MFCC are obtained with its derivatives. The 39-dimensional vectors are fed to the supervised SVM for the classification of voice regions [7].

3.5. *Method-5 (m5)*

The nonlinear dynamic methods are capable of detecting abnormal vocal cord behavior, used in clinical practice investigation, voice classification, as suggested by Titze. Such non-linear methods are adopted in the present work. Dynamic parameters with positions and velocities are described as phase space. In phase space with time evolution, the vibrations produced by a complex system such as the vocal folds are being shown as a projected path. Periodic vibrations generate a closed course and an irregular trajectory depicts aperiodic vibrations [8]. Nonlinear features like MI, FNN, CD, CE2 (information dimension), RE-1, RE-2 and LTE are extracted. Table 2, shows the further experimental investigation of methods m_6 to m_8 with a combination of linear and nonlinear features. The respective linear and nonlinear feature extraction has been explained in section 3.

4. Materials

In the present experimental study, the NOIZEUS database containing 30 IEEE continuous speech sentences corrupted by real-world car or babble noises at varying levels of SNRs [5] is used. The remaining two databases, SVD and Laryngeal cancer databases are focused on the study of sustained vowels. SVD is an online, freely available voice disorder database containing a set of voice recordings [19] with 16-bit resolution at 50 kHz. In the present work, 40 sustained vowels [a,i,u] of different larynx pathologies (vocal fold paralysis-14, Laryngeal carcinoma-14, Chronical laryngitis-8, hypo-pharyngeal carcinoma-2, vocal fold nodules-2) are considered. For the laryngeal cancer database 50 pathological voices are recorded (sustained vowels /a/) at 44.1 kHz, 16-bit resolution at the Sri Siddivinayak Cancer Hospital (SSGCH), Miraj and NDMCH, Barshi, India with the consent of each case. In advance ethical committee approval was taken. These cases have covered the adjacent regions of the vocal cords. Because the authors believe that, the vocal cords have more dynamics as far as the voice of subject having Laryngeal pathology.

5. Experimental Setup

The present work is split up into two main parts. The first part deals with the experimental investigation of the first five types of VADs using SVM as shown in Table 1. The first four methods are based on the linearity assumption. In the m1 method (4xnumber of frames) vector per frame and in the method m_2, a (39xnumber of frames) vector per frame is fed to the SVM. In the method m_3, a 39-dimensional feature vector per frame and in m_4, 39-dimensional vector per frame is used as input to SVM with simple energy-based voting. In the fifth method m_5, six nonlinear features are extracted per frame as explained in section 3 and is fed to SVM with energy-based voting. The second part deals with the experimental investigation of methods (m_6, m_7, m_8) involving the combination of linear and nonlinear features using SVM as shown in Table 2. The dimensions of feature vectors fed to SVM using these methods m_6 to m_8 are 45, 49 and 10 respectively. The performance of each of the methods (m_1 to m_8) using the present framework is evaluated by following Eqn's,

$$\text{Accuracy}=((\text{TP}+\text{TN})/\text{total}) * 100 \quad (11)$$

$$\text{Sensitivity(TPrate or Recall)}(\text{TP}./(\text{TP}+\text{FN})) * 100 \quad (12)$$

$$\text{Specificity}=\text{TN}/(\text{TN}+\text{FP}) * 100 \quad (13)$$

$$\text{FP}_{\text{rate}} = (\text{FP}./(\text{FP}+\text{TN})) * 100 \quad (14)$$

$$\text{MissClassification}_{\text{rate}} = ((\text{FP}+\text{FN})/\text{total}) * 100 \quad (15)$$

$$\text{Precision}=(\text{TP}/(\text{TP}+\text{FP})) * 100 \quad (16)$$

$$\text{AUC}=0.5x(\text{Specificity}+\text{Sensitivity}) \quad (17)$$

$$\text{F Score}=2.\text{Precision}.\text{Recall}/(\text{Precision}+\text{Recall}) \quad (18)$$

The parameters are calculated from the confusion matrix, which is extracted using SVM. True positive (TP) indicates speech presence and true negative (TN) denotes the absence of speech.

6. Results and Discussion

The results are discussed as follows;

6.1. *Results corresponding to NOIZEUS database*

Fig.3(a) shows the graph of SNR levels versus area under curve (AUC), which shows the discrimination capability between the speech and non-speech segments across all methods. The AUC values rise from 0.90 to 0.96 with SNR levels varying from 0 dB to 15 dB. It shows that the classification rates are better for the methods m_6, m_7 and m_8, which are using a combination of linear and nonlinear features as compared with the other methods. The FP-rate is highly reduced (9.9 % to 6.4%) in these methods as shown in Fig.3 (b). The methods m_2 and m_3, based on gammatone features have shown a larger FP- rate (47.17% to 50.49%) with lower AUC values (0.31 to 0.77). Moreover, method m_5, which is purely based on nonlinear parameters, has shown FP-rate of (22.5% to 48.7%) with lower AUC values (0.42 to 0.76). It is clear from the Fig.3 (b) that, the FP-rate has been decreasing with subsequent increase in AUC values from 0 dB to 15 dB as far as all the methods are concerned.

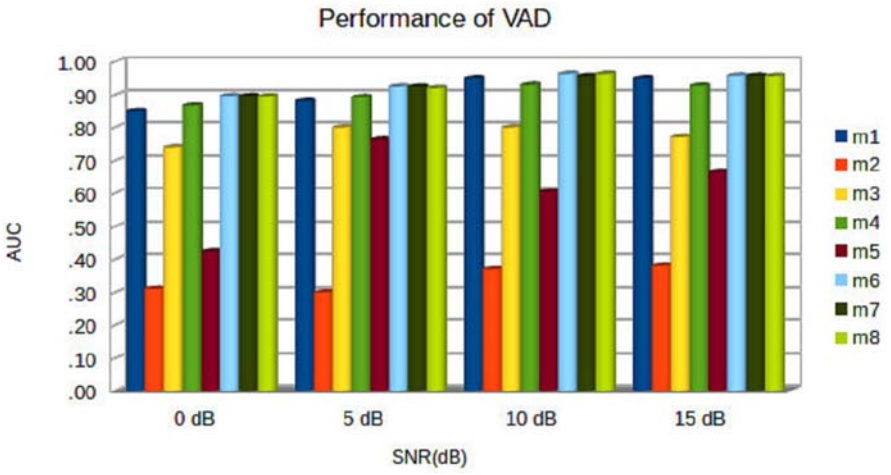


Fig. 3.(a) Performance of the proposed VAD for all the methods using the NOIZEUS database at varying SNR levels in terms of area under the curve (AUC)

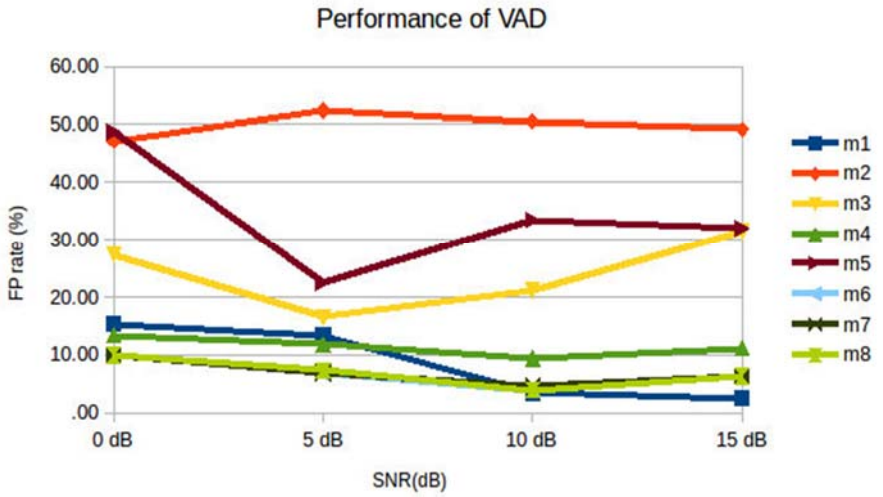


Fig. 3(b) Variation of Percentage of False Positive Rate for all the methods using the NOIZEUS database at varying SNR levels

Table 3, shows the experimental results revealing the success of eight methods to varying SNR levels (0 dB to 15 dB) of babble noise in terms of accuracy, specificity and F-score. The methods based on linearity assumption like m_2, m_3, and m_5 have shown poor accuracy F-score with a subsequent gradual increase from 0 dB to 15 db. The methods m_5, m_6, and m_7 using the combination of linear and nonlinear features have shown better accuracy. The method m_6 involving bio-inspired MFCC and nonlinear features has shown an accuracy of 89%-92% at lower SNR levels of 0-5 dB and depicting a sudden rise in performance around 96%, at higher SNR levels from 10-15 dB. The method m_6 has shown a larger average accuracy of 93.5% compared to m_7 and m_8. This shows that the combination of nonlinear parameters with bio-inspired MFCC features is more robust in VAD as compared with combination to other time domain based linear features. However, as m_1 has a lowest F-score and m_7 has got highest compared to all methods. Hence, hybrid methods appear to be better for robust VAD with supervised SVM.

Methods	0 dB			5 dB			10 dB			15 dB		
	Accuracy	Specificity	F Score	Accuracy	Specificity	F Score	Accuracy	Specificity	F Score	Accuracy	Specificity	F Score
m1	89.62	84.62	0.90	88.54	86.59	0.92	93.82	96.42	0.96	93.77	97.42	0.95
m2	49.15	32.83	0.83	43.60	37.59	0.88	47.82	49.51	0.90	49.61	50.81	0.91
m3	76.31	72.65	0.84	76.58	83.44	0,84	78.81	78.86	0.85	78.69	68.60	0.83
m4	86.56	86.57	0.90	89.57	88.19	0.92	93.64	90.65	0.95	93.60	88.80	0.95
m5	51.03	51.32	0.88	78.05	77.51	0.79	68.89	66.54	0.92	72.98	68.01	0.91
m6	89.32	90.10	0.92	92.03	93.13	0.94	96.27	96.05	0.97	96.40	93.66	0.97
m7	89.27	90.10	0.92	92.03	93.14	0.94	95.70	95.35	0.97	96.40	93.66	0.97
m8	89.16	90.07	0.92	91.62	92.66	0.94	96.27	96.05	0.97	96.34	93.66	0.97

Table 3. Summary of the results obtained by using proposed VAD for all the methods with NOIZEUS database at varying SNR levels

The Fig.4 depicts the SVM prediction ability in all the methods across varying SNR levels. It is found that the difference between ideal and predicted labels is less in all the methods except in m_2, m_3, m_4, and m_5 methods. In summary, as far as continuous speech is concerned, the methods m_1, m_5, and m_6 methods appear to be more robust in VAD using SVM. The main reasons attributed to the success of these methods m_5 and m_6 are the involvement of the bio-inspired MFCC and non-linear features, which are able to explore the dynamics of the vocal folds.

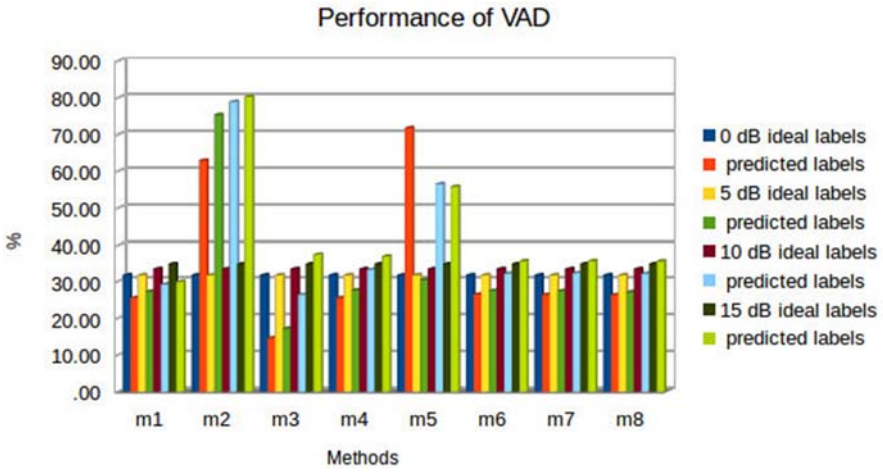


Fig. 4. Comparison of Rate of Prediction (%) across all methods using the NOIZEUS database at varying SNR levels

6.2. *Results corresponding to Saarbruecken Voice Database (SVD) and Laryngeal cancer database se*

Here, 40 sustained vowels /a/ from SVD and 50 from the third database are together used in the experimental investigation. The results are summarized as shown in Table 4. The accuracy of more than 80% is shown by the methods m_1, m_3, m_4, and m_6. Even though the rate of ideal labels is 63.99%, the rate of predicted labels is found to be nearer to the ideal labels using m_5, m_7, and m_8 methods. However, the method m_4 shows better accuracy of 89.99% with better TP-rate, specificity, precision, AUC, and F-score. But, the method m_4 has a lower prediction rate as compared with m_5. The lower or higher prediction rates are attributed to the recording time interval of vowels from 1 second to 3 seconds. Hence, FP-rate is higher in the case of methods m_5 to m_8 involving nonlinear parameters which, are more sensitive to noise or non-speech portions in the voice. The main reasons attributed to this performance as far as vowels are concerned is, the significant involvement of phase by nonlinear parameters. In sustained vowels, the methods involving hybrid parameters have shown better SVM performance.

Methods	Ideal lables	Predicted labels	Accuracy	Sensitivity	FP Rate	Specificity	Precision	AUC	F Score
m1	63.99	80.00	81.78	78.16	17.77	82.23	80.00	0.80	0.99
m2	63.99	30.91	47.54	28.97	11.29	88.71	28.83	0.59	0.96
m3	63.99	52.38	90.34	52.35	9.17	90.83	99.36	0.64	0.95
m4	63.99	57.31	92.61	66.67	10.60	89.40	95.46	0.65	0.98
m5	63.99	66.89	67.65	50.00	25.17	74.83	64.49	0.44	0.98
m6	63.99	54.24	99.94	57.14	11.59	88.41	94.92	0.64	0.96
m7	63.99	67.92	80.97	40.00	24.80	75.20	58.57	0.44	0.98
m8	63.99	68.21	80.74	40.00	24.80	75.20	58.57	0.44	0.98

Table 4 Summary of the results obtained by using proposed VAD for all the methods with SVD and third database

7. Conclusions

The present framework for reliable voice activity detection is focusing on sustained vowels and continuous speech with babble noise. The framework is supported by a two-step noise reduction method of the Wiener and SVM classification. The implementation of the entire framework using an open-source Linux platform and GNU Octave is cost-effective. Experimental investigations are carried out using NOIZEUS continuous speech corpus with babble noise. The results revealed that the methods involving linear and nonlinear features, m_1, m_5, and m_6 appear to be more robust in VAD using SVM with an accuracy around 93%. For sustained vowels, the laryngeal pathologies data from SVD and laryngeal cancer data depicted that, the time-domain based VAD methods having accuracy of around 80% with better TP rate, specificity, precision, AUC and F-score. SVM prediction rate is found to be better in the methods m6 (ideal 63.9% and predicted 67.9%) along with much reduced FP-rate of 11.59% and a bit higher FP-rate in method 7. In summary, the combination of biologically inspired features like MFCC and nonlinear features has proved to be the emphasis in the reliable VAD and further assessment of signal to noise ratio, which is useful in speaker verification. Because the dynamic behaviour of vocal folds is more prominently revealed by the consideration of phase in nonlinear features in the VAD methods m_6 and m_7. This VAD forms an integral part of the pre-processing stage of speech involving continuous and sustained vowels affected with babble noise in case of pathological voice analysis and detection. There is more scope in the development of VAD by an in-depth analysis of nonlinear methods and higher-order machine learning methods with the involvement of phase and magnitude.

References

[1] R. Amir Hossein Poorjam, Max A. Little, Jesper Rindom Jensen and Mads Græsboll Christensen, (2018),"A Supervised Approach to Global Signal-to-Noise Ratio Estimation for Whispered and Pathological Voices", 978-1-5386-4658-8/18/2018 IEEE, ICASSP.Vol. 1, No. 1, pp. 123-456, 2009.

[2] Cyril Plapous, Claude Marro, and Pascal Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement", *IEEE Transactions on Audio, Speech and Language Processing.*2006.

[3] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, "*A Practical Guide to Support Vector Classification*, Department of Computer Science", National Taiwan University, Taipei 106, Taiwan, May 2006.

[4] Carlos M. Travieso, Jesús B. Alonso, J.R. Orozco-Arroyave, J.F. Vargas-Bonilla, E. Nöth, Antonio G. Ravelo-García, "Detection of different voice diseases based on the nonlinear Characterization of speech signals", *Expert Systems with Applications* (pp. 184–195), 2017.

[5] Dang Minh Cong, (2015), "*Noise Reduction In Speech Enhancement By Spectral Subtraction With Scalar Kalman Filter"*, Ha Noi – 2015.

[6] Broder, Gerard Chollet, Anna Esposito, Marcos Faundez-Zanuy, Maria Marinaro, "*Non Linear Speech Modeling and Applications: Advanced Lectures, and revised selected papers"*, Springer, May, 2005.

[7] Chakrabarti, Hamzeh Ghasemzadeh, Mehdi Tajik Khass, Meisam Khalil Arjmandi, Mohammad Pooyan, "Detection of vocal disorders based on phase space parameters and Lyapunov spectrum", *Biomedical Signal Processing and Control*, *Elsevier*-2015.

[8] Lasonas Kokkinos, Student Member, Petros Maragos, "Nonlinear Speech Analysis Using Models for Chaotic Systems", *IEEE Transactions On Speech and Audio Processing*, Vol. 13, No. 6, Nov, 2005.

[9] In-Chul Yoo, Hyeontaek Lim, Dongsuk Yook, "Formant-Based Robust Voice Activity Detection", *IEEE/ACM Transactions On Audio, Speech, and Language Processing*, Vol. 23, No. 12, Dec, 2015.

[10] Jack J. Jiang, Yu Zhang, and Clancy McGilligan, "Chaos in Voice, From Modeling to Measurement", *Journal of Voice*, Vol. 20, No. 1, pp. 2–17, 2006.

[11] Jun Tae Kim, Sung Hoon Jung, Kwang-Hyun Cho, "Efficient harmonic peak detection of Vowel sounds for enhanced voice activity detection", *IET Signal Process*, Vol. 12 Iss. 8, pp. 975- 982, 2018

[12] Lavneet Singh, Girija Chetty and Savleen Singh, "A Novel Algorithm Using MFCC and ERB Gammatone Filters in Speech Recognition", *Journal of Information Systems and Communication*, ISSN: 0976-8742 & E-ISSN: 0976-8750, Volume 3, Issue 1, pp. -358-364, 2012.

[13] M. H. Moattar and M. M. Homayounpour,"A Simple But Efficient Real-Time Voice Activity Detection Algorithm", *17th European Signal Processing Conference* (EUSIPCO 2009) Glasgow, Scotland, Aug, 2009.

[14] Md. Moinuddin, Arunkumar N. Kanthi, "Speaker Identification based on GFCC using GMM"", *International Journal of Innovative Research in Advanced Engineering* ISSN: 2349-2163, Volume 1, Issue 8, 2014.

[15] Nitin N Lokhande, Navnath S Nehe, Navnath S Nehe, "Voice Activity Detection Algorithm for Speech Recognition Applications", *International Conference on Computational Intelligence* (ICCIA), 2011.

[16] Pham Chau Khoa, "*Noise Robust Voice Activity Detection"*, School of Computer Engineering, 2012.

[17] Rainer Hegger, Holger Kantz, Thomas Schreiber, *"Practical implementation of nonlinear time series methods: The TISEAN package"*.

[18] Simon Graf, Tobias Herbig, Markus Buck and Gerhard Schmidt, "Features for voice activity detection: a comparative analysis", *EURASIP Journal on Advances in Signal Processing* 2015:91, DOI 10.1186/s13634-015-0277-z.

[19] W. J. Barry and M. Putzer, "*Saarbrucken Voice Database",* Institute of Phonetics, Univ. of Saarland, http://www.stimmdatenbank.coli.uni-saarland.de/.

[20] Wilson Burgos, "*Gammatone and MFCC Features In Speaker Recognition"*, Melbourne, Florida-2014.

[21] W. Q. Ong and A. W. C. Tan," Robust Voice Activity Detection Using Gammatone Filtering and Entropy", *International Conference on Robotics, Automation and Sciences* (ICORAS)-2016.

[22] Yan Zhang, Zhen-min Tang,Yan-ping Li,Yang Luo, "A Hierarchical Framework Approach for Voice Activity Detection and Speech Enhancement, Hindawi Publishing Corporation, *Scientific World Journal*, Volume 2014.

[23] Youngmoon Jung, Younggwan Kim, Hyungjun Lim, Hoirin Kim,"Linear-Scale Filterbank for Deep Neural Network-Based Voice Activity Detection", *Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique* (O-COCOSDA) Seoul, Korea, 2017.

[24] Zulfiqar Ali, Muhammad Talha, "Innovative Method for Unsupervised Voice Activity Detection and Classification of Audio Segments", *IEEE Access*, Volume 6, 2018.