I.J. Image, Graphics and Signal Processing, **2017, 8, 50-58**
Published Online August 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2017.08.06

Robust Voice Activity Detection Algorithm based on Long Term Dominant Frequency and Spectral Flatness Measure

Naorem Karline Singh and Yambem Jina Chanu
Department of Computer Science and Engineering, National Institute of Technology Manipur, 795004, India
Email: {naoremkarline, jina.yambem}@gmail.com

Received: 18 April 2017; Accepted: 13 May 2017; Published: 08 August 2017

Abstract—In this paper, a robust voice activity detection algorithm based on a long-term metric using dominant frequency and spectral flatness measure is proposed. The propose algorithm makes use of the discriminating power of both features to derive the decision rule. This method reduces the average number of speech detection errors. We evaluate its performance using 15 additive noises at different SNRs (-10 dB to 10 dB) and compared with some of the most recent standard algorithms. Experiments show that our propose algorithm achieves the best performance in terms of accuracy rate average over all SNRs and noises.

***Index Terms*—**Voice activity detection, dominant frequency component, spectral flatness measure.

I. Introduction

Voice activity detection (VAD), is an essential pre-processing step in many speech and audio processing applications such as automatic speech recognition [1], speaker diarization [2] and speaker identification systems [3]. VAD is often referred to the process of classifying speech and non-speech regions in an audio signal. Non-speech regions may be silence, noise, music or other complex acoustic signal such as recording in streets, train stations, etc. It is mainly used to achieve high recognition rate or system accuracy by removing insignificant parts while processing the signal. It is also used in real time communication systems [4] and speech encoder [5] to attain high compression rate and low transmission rate.

VAD can be classified according to features it uses or the nature of implementing its decision mechanism i.e. supervised or unsupervised [6]. Earlier VAD techniques are based on time domain and low dimensional features such as energy [7], zero crossing rate [8], line spectral frequency [7] and autocorrelation [9]. Frequency domain [10, 11] VAD algorithm tends to perform better than time domain algorithm. Most of these VADs operate on short-term window (frame) and their discriminative power drops when SNR fall below 10dB. Over the past few decades many new complex features are introduced exploiting the spectral properties of speech and non-speech regions in an audio stream. In contrast to the use of short-term frame level, Ramirez et al. [12] propose the use of long term spectral divergence (LTSD) between speech and non-speech, which require average noise spectrum magnitude which is not practically available. Experimental results show that VAD decision taken over long term analysis window is more accurate than short-term window for noisy environments [12–14]. Fukuda et al. [13] propose the long-term dynamic feature for VAD using cepstrum of neighbor frames. Ghosh et al. [14] propose long term signal variability (LTSV) based VAD which measures the sample variance of long-term sub-band entropies. LTSV shows great improvement in both stationary and non-stationary noise conditions, but its discrimination power drops when SNR is higher than 5dB. Moreover, Yanna ma et al. [15] propose long term spectral flatness measure (LSFM) based VAD, which employs a low-variance spectrum estimate and an adaptive threshold. LSFM-based VAD performs well for most noise types even in low SNR but, fails for some specific noises. Recently some VAD based on artificial neural networks have also been introduced [16, 17] using robust acoustic features and most of them are unsupervised learning. Statistical model-based VAD is also becoming popular, classifier are mainly based on Gaussian Mixture model (GMM) [6] and Support Vector Machine (SVM) [18]. Most of these methods mention above assume noise to be stationary for a certain period, which made them sensitive to change in SNR of the observed signal. SNR estimation to improve VAD robustness is a difficult task for non-stationary noises. Therefore, design of VAD algorithm which can work in very low SNR is necessary.

Spectrum of speech regions have non-uniform power and thus have low spectral flatness whereas noise regions exhibit high spectral flatness as shown in fig. 1(a) and fig. 2(a). Spectral flatness using long-term window perform well for SNR above 0 dB, but under low SNR (below 0 dB) it tends to saturate its discriminating power with increase in speech detection error.

In this paper, we have proposed a new improve VAD algorithm based on long term dominant frequency and spectral flatness measure. To reduce the effect of misclassification of speech frame in low SNR, dominant

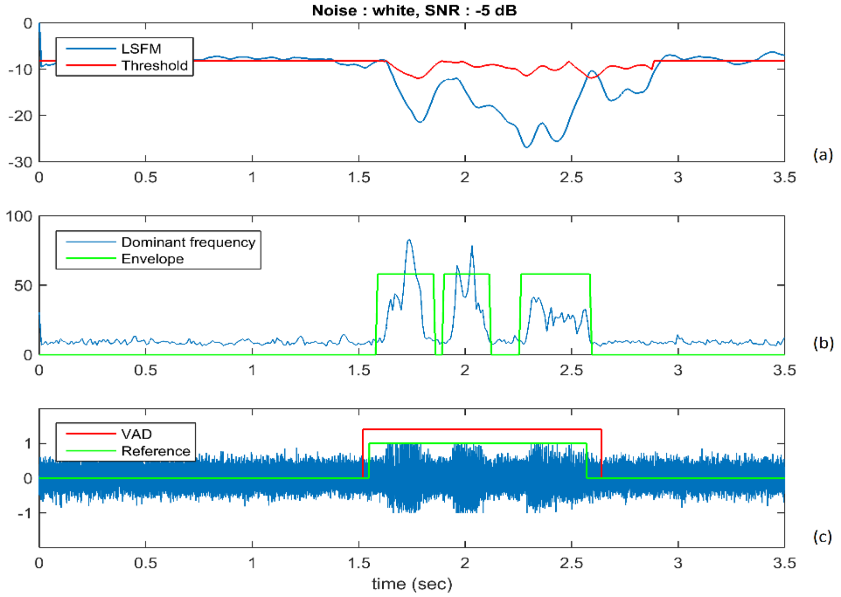

Fig.1. Illustrative example of proposed VAD algorithm on a randomly chosen clean-speech sentence from CTSR [20] noisy speech database test set, with white noise added at -5 dB SNR: (a) shows LSFM value and adaptive threshold; (b) shows the dominant frequency component and spectral frequency envelope of speech region; and (c) shows the VAD output and actual speech reference label.

frequency component of a speech signal is used along with LSFM feature. Dominant frequency of a speech signal give better discrimination than LSFM in terms of speech and non-speech boundary. We have verified the usefulness of the combined feature by analyzing the discriminative power under various noise types and SNR conditions.

The organization of the paper is as follows: section II describes dominant frequency and spectral flatness measure based features and their discriminative power. In section III, we explain our proposed algorithm. Section IV, describes our implementation detail and datasets used in this paper and our evaluation results. Finally, conclusion is given in section V.

II. Dominant Frequency Component and Spectral Flatness Measure

Selection of features which are robust against various types of noises will lead to increase in discriminating power of the system. Dominant frequency component and spectral flatness measure are two such features which have high noise robustness.

A. Dominant frequency component

In a noisy environment where most of the speech region is corrupted by noise, it is desirable to enhance the speech region that has more energy or are dominant. Dominant frequency component of the speech sample is computed by finding the frequency corresponding to the maximum value of the spectrum magnitude. There are many methods for finding the dominant frequencies which are discussed in [19] and FFT seems to be the best method for estimating dominant frequency of the signal. Steps involved in computing dominant frequency are:

1) Uniformly segment the recorded noisy signal using a hamming window of 20ms frame size and a frame shift of 10ms.
2) Apply N-point FFT for each frame to compute power spectral density.
3) Find peaks in each frame. Frequency of the sample which have the highest peak correspond to dominant frequency component of that particular frame.

Setting an appropriate fixed threshold to classify speech regions will work for stationary noises. But in real life most of the noises are non-stationary, so instead of setting a fixed threshold we develop a new method which can work well for most noise cases. From fig. 1(b), fig. 2(b), fig. 3(b) and fig. 5(b) we can see that speech region have higher peaks as compare to non-speech region and another point is that they have larger envelope, which is an important factor to make the classifier. Steps for creating spectral frequency envelope are:

1) Find the initial threshold which is the average of the first 100 dominant frequency component.
2) Find the starting and ending frame of each envelope. If the dominant frequency of a frame is greater than initial threshold then it is set as starting frame. The last frame of the successive frames whose dominant frequency is greater than initial threshold is set as ending frame of that envelope.
3) Next find the average envelope size (number of frames) from the initial 1.5s silence region.
4) Remove all those envelopes whose envelope size is

less than twice the average envelope size.

LSFM feature tends to misclassify speech frames while passing from speech to non-speech region. This is due to the spectral information it carries from speech region leading to non-uniform spectral power. This error can be reduced to some extent by marking boundary of speech region using dominant frequency.

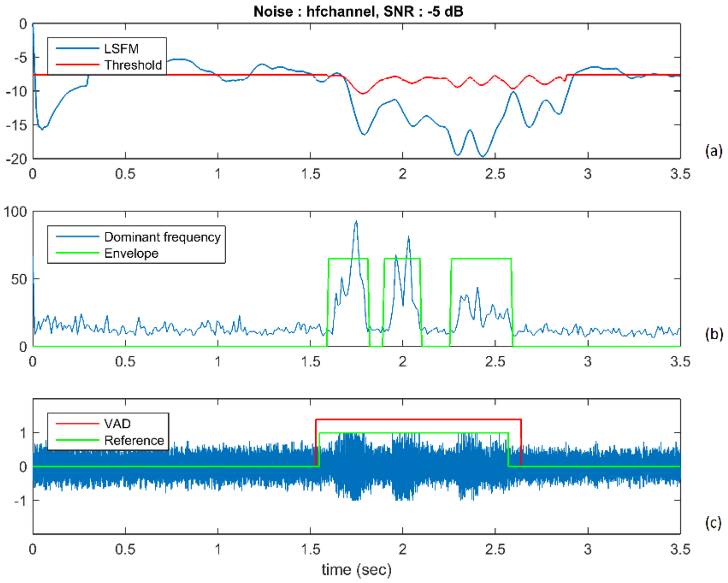

Fig.2. Illustrative example of proposed VAD algorithm on a randomly chosen clean-speech sentence from CTSR [20] noisy speech database test set, with high frequency channel noise added at -5 dB SNR: (a) shows LSFM value and adaptive threshold; (b) shows the dominant frequency component and spectral frequency envelope of speech region; and (c) shows the VAD output and actual speech reference label.

B. Long term spectral flatness measure

LSFM feature, $L_x(m)$ of a given signal x at the m^{th} frame is given by the ratio of geometric and arithmetic mean of the power spectrum. Value of $L_x(m)$ lies in the range $(-\infty,0]$ with the maximum value acquire when the geometric mean is equal to the arithmetic mean .

$$L_x(m) = \sum_k log_{10} \frac{GM(m,w_k)}{AM(m,w_k)} \qquad (1)$$

Where $GM(m,w_k)$ is the geometric mean and $AM(m,w_k)$ is the arithmetic mean of the power spectrum $S(n,w_k)$.

$$GM(m,w_k) = \sqrt[R]{\prod_{n=m-R+1}^{m} S(n,w_k)} \qquad (2)$$

$$AM(m,w_k) = \frac{1}{R}\prod_{n=m-R+1}^{m} S(n,w_k) \qquad (3)$$

Where R is the number of last frames used to compute LSFM metric and $S(n,w_k)$ is the short-time spectrum of M consecutive frames.

$$S(n,w_k) = \frac{1}{M}\sum_{p=n-M+1}^{n} |X(p,w_k)|^2 \qquad (4)$$

$$X(p,w_k) = \sum_{l=(p-1)N_{sh}+1}^{N_w+(p-1)N_{sh}} w(l-(p-1)N_{sh}-1)x(l)e^{-jw_k l} \qquad (5)$$

Where $X(p,w_k)$ is the short-time Fourier transform coefficient at frequency w_k of the p^{th} frame. $w(i)$ is the short-time Hann window , and $i \in [0,N_w)$. N_w is the frame length and N_{sh} is the frame shift duration in terms of samples.

1) Selection of Frequency Range, w_k

For better discrimination between speech and non-speech region, choosing a frequency range for intelligible speech is necessary. Since speech is a low pass and non-stationary signal, 500 Hz to 4 kHz is necessary for speech intelligibility. The frequency range require in computing LSFM feature is given below

$$k = N_{DFT}\left(\frac{4000-500}{f_s}\right) \qquad (6)$$

Where N_{DFT} is the order of discrete fourier transform coefficient used in estimating the spectral estimate and f_s is the sampling frequency.

2) Threshold Estimation

We assume that the initial 1.5s of our input signal is always silence region. From this region 100 realizations of LSFM features is stored in buffer, ψ_L. Then, the initial threshold γ_L is determined as follow:

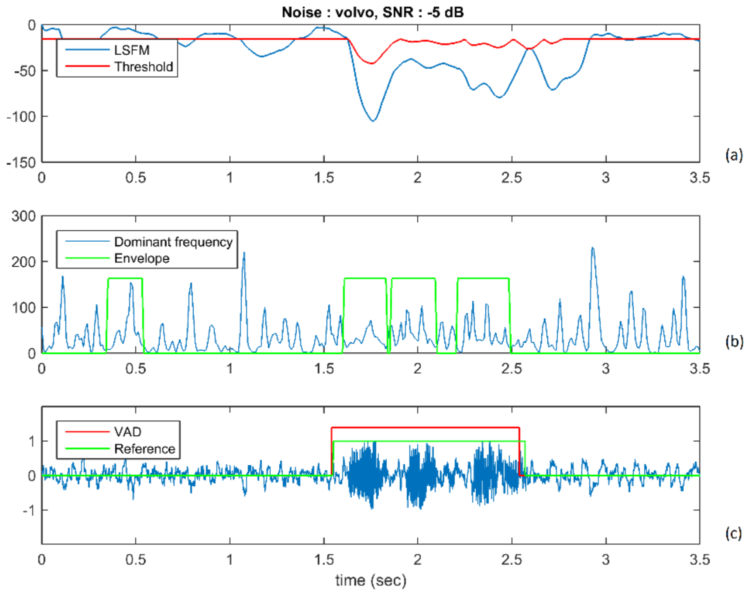

Fig.3. Illustrative example of proposed VAD algorithm on a randomly chosen clean-speech sentence from CTSR [20] noisy speech database test set, with volvo noise added at -5 dB SNR: (a) shows LSFM value and adaptive threshold; (b) shows the dominant frequency component and spectral frequency envelope of speech region; and (c) shows the VAD output and actual speech reference label.

$$\gamma_L = mean(\psi_L) \tag{7}$$

Since, a fixed threshold won't work for all types of noises therefore, for every detection of speech frame while determining the initial decision, we update the threshold. The new threshold at frame m^{th} is given by

$$\gamma_L(m) = \sigma_L + \gamma_L \tag{8}$$

Where σ_L is the variance (standard deviation) of the L_x values for the last 100 frames.

III. The Proposed Algorithm

A flow-chart diagram of the proposed VAD algorithm is shown in fig. 4. Steps involved in this algorithm can be described as follows. First the input noisy signal is pre-processed using a simple spectral subtraction technique [21] to filter out the background stationary noise. After spectral subtraction, the input signal is segmented into frames of 20ms in length and a frame-shift of 10ms. Dominant frequency component $D_x(m)$ of each frame is calculated using steps described in section II-A and spectral envelopes are also estimated. Then we follow the same procedure for LSFM feature $L_x(m)$ computation as stated in Yanna Ma *et al.* [15]. The power spectrum of the segmented signal is estimated using Welch-Bartlett method since it is better than periodogram [22].

A. Decision Rule

The initial decision about whether there is a speech frame is determined by using the previous R frames. Frame m^{th} is said to be a speech frame if the value of $L_x(m)$ is greater than its corresponding threshold γ_L and that frame is within a spectral envelope. The initial decision V_INL at m^{th} frame is set to 1 if there is any speech frame in the previous R frames otherwise V_INL is set to 0. For smoothing the initial decision we apply the voting scheme from [14] to obtain VAD decision at every 10ms. The target 10ms is taken as a speech frame if there is 80 percent or more speech frames in the previous R initial decision.

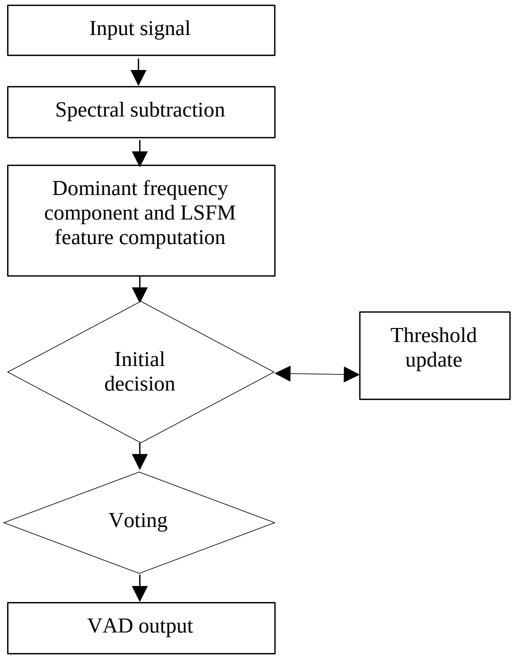

Fig.4. Flow-chart diagram of the proposed VAD algorithm.

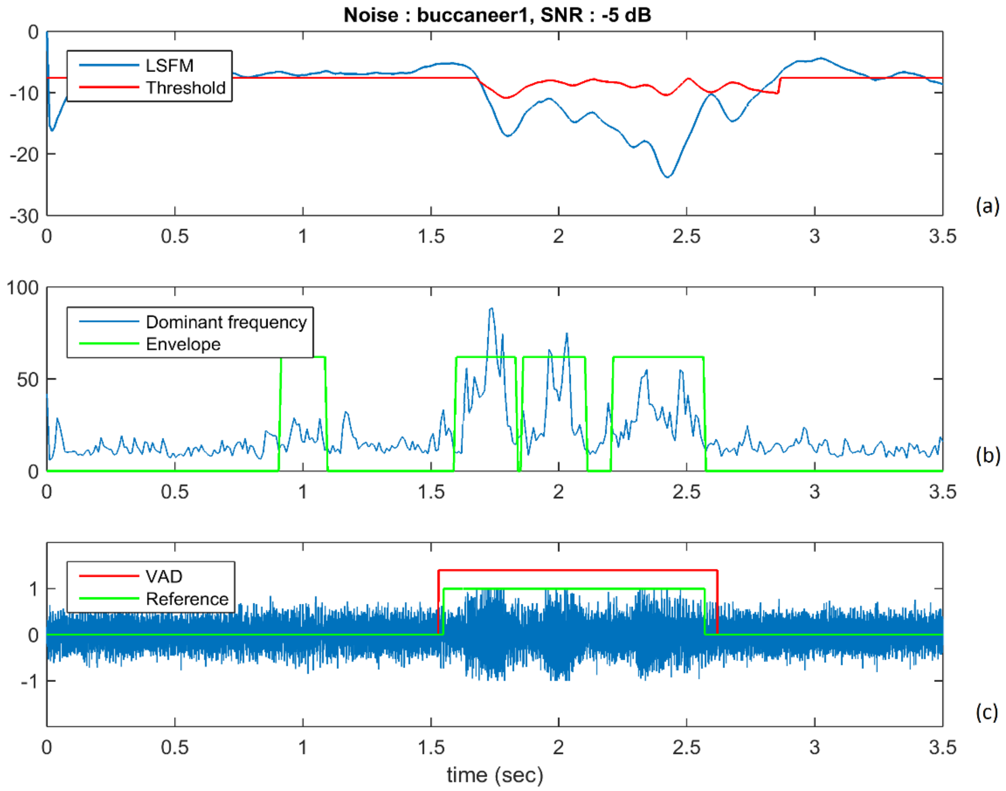

Fig.4. Illustrative example of proposed VAD algorithm on a randomly chosen clean-speech sentence from CTSR [20] noisy speech database test set, with buccaneer noise added at -5 dB SNR: (a) shows LSFM value and adaptive threshold; (b) shows the dominant frequency component and spectral frequency envelope of speech region; and (c) shows the VAD output and actual speech reference label.

B. *Selection of R and M*

R and M are parameters used for computing the LSFM feature and R is also used for determining the initial decision. Proper selection of this two parameter will increase discriminating power of speech and non-speech and hence a better VAD. We evaluate our proposed algorithm following the method mentioned in [15] using the Edinburgh corpus database and NOISEX92 database (in section IV). Experimentally we also found that the values for R = 30 and M = 10 are same as in [15].

IV. Experiments And Results

In order to analyze the performance of the proposed VAD, clean speech and different noises datasets are required. Two datasets are needed for testing and training the proposed system. Steps involved in the data preparation and experimental setup are described in sub-section A, evaluation metric in sub-section B and finally the comparison of various VAD's is given in sub-section C.

A. *Data and Experimental Setup*

To evaluate the proposed method, clean speech test set dataset [20] from University of Edinburgh, Centre for Speech Technology Research is used. Test set contains clean speech of 400 sentences each spoken by 2 native English speakers. Each sentence is no longer than 10 s and on average 80 % of each sentence are labelled as speech. Hence, to make it comparable to real conversational speech, randomly chosen sentences are concatenated by adding 1.5s silence at the beginning, ending and junctions of the utterances. Two dataset, training and testing are constructed using the above process and size of each dataset is around 500s. And for evaluation purpose, reference speech labels are created by manually hand labelling the speech and non-speech regions using the software wavesurfer [23]. All 15 noises from the NOISEX92 [24] database are added to both the testing and training set at 5 different SNR (-10 dB,-5 dB, 0 dB, 5 dB, 10 dB). Resulting two dataset are then used to evaluate the system parameters and evaluation purpose respectively. List of all the noises are given below:

- Two types of factory floor noises (near car-production hall and near plate cutting and electrical welding equipment)
- Three types of Cockpit noises (Buccaneer jet travelling at 450 knots , 190 knots and F-16 jet at 500 knots)
- Two types of engine noises (Destroyer engine room noise and engine operation room background noise)

Copyright © 2017 MECS

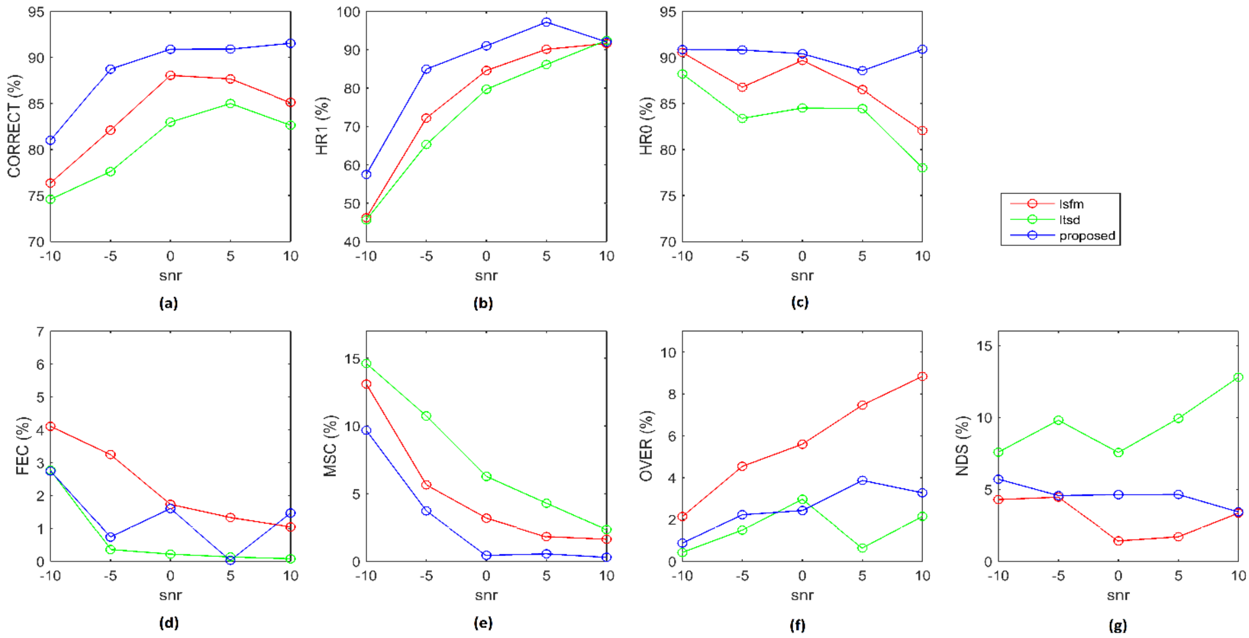

Fig.6. Comparison of three VAD algorithms averaged over 15 noises for five SNR levels in terms of accuracy rate - (a) CORRECT (b) HR1 (c) HR0 and error rate – (d) FEC (e) MSC (f) OVER (g) NDS.

- Two types of military vehicle noises (M109 Tank noise moving at 30km/h and leopard 1 vehicle moving at 70 km/h)
- Speech babble noise (100 people speaking in canteen)
- High frequency radio channel noise
- Pink noise
- White noise
- Machinegun noise (.50 caliber gun fired repeatedly)
- Vehicle interior noise (Volvo 340 moving at 120 km/h)

B. *Evaluation Metric*

For evaluating the performance of the proposed VAD algorithm we follow the objective evaluation methods [4], where labels obtained by the VAD is compared against true reference labels. Objective evaluation can be done in two ways - accuracy rate and error rate. Parameters used for the performance evaluation are as follows:

1) CORRECT : correct decision made by the VAD algorithms
2) Speech hit rate (HR1): speech frames detected correctly among all speech frames.
3) Non-speech hit rate (HR0): non-speech frames detected correctly among all non-speech frames.
4) Front end clipping (FEC): speech misclassified as non-speech in passing from non-speech to speech region.
5) Mid-section clipping (MSC): speech misclassified as non-speech in a speech region.
6) Carry over (OVER): non-speech misclassified as speech in passing from speech region to non-speech.
7) Noise detected as speech (NDS): non-speech misclassified as speech within a non-speech region.

Among these seven parameters CORRECT, HR1, HR0 gives the correct decision made by the VAD algorithm which is the accuracy rate of the system. These parameters should be maximized in order to achieve best system performance. And the remaining four parameters- FEC, MSC, OVER and NDS gives the false detection (error rate) made by the system. These four parameters need to be minimized since they lead to poor performance of system. Among these four parameters, MSC should be taken utmost care since, its increase will lead to miss of actual speech region. To illustrate the performance of our proposed VAD, two standard VAD algorithms are chosen for comparison. They are LTSD [13] and LSFM [16]. Both of them are implemented in matlab as according to their papers. The order of LTSD is 6. And for LSFM the long term window parameter are set as (R = 30 and M = 10).

C. *Evaluation Results*

Comparisons with other standard VAD's is performed in two ways. Firstly in terms of average accuracy and error rate for all 15 noises at five different SNR levels and secondly by averaging over five SNR levels for all 15 noises.

Figure 6 shows the average accuracy and error rate of three evaluated algorithms for all 15 noises. Here the first row provides the three accuracy rate metric- (a)

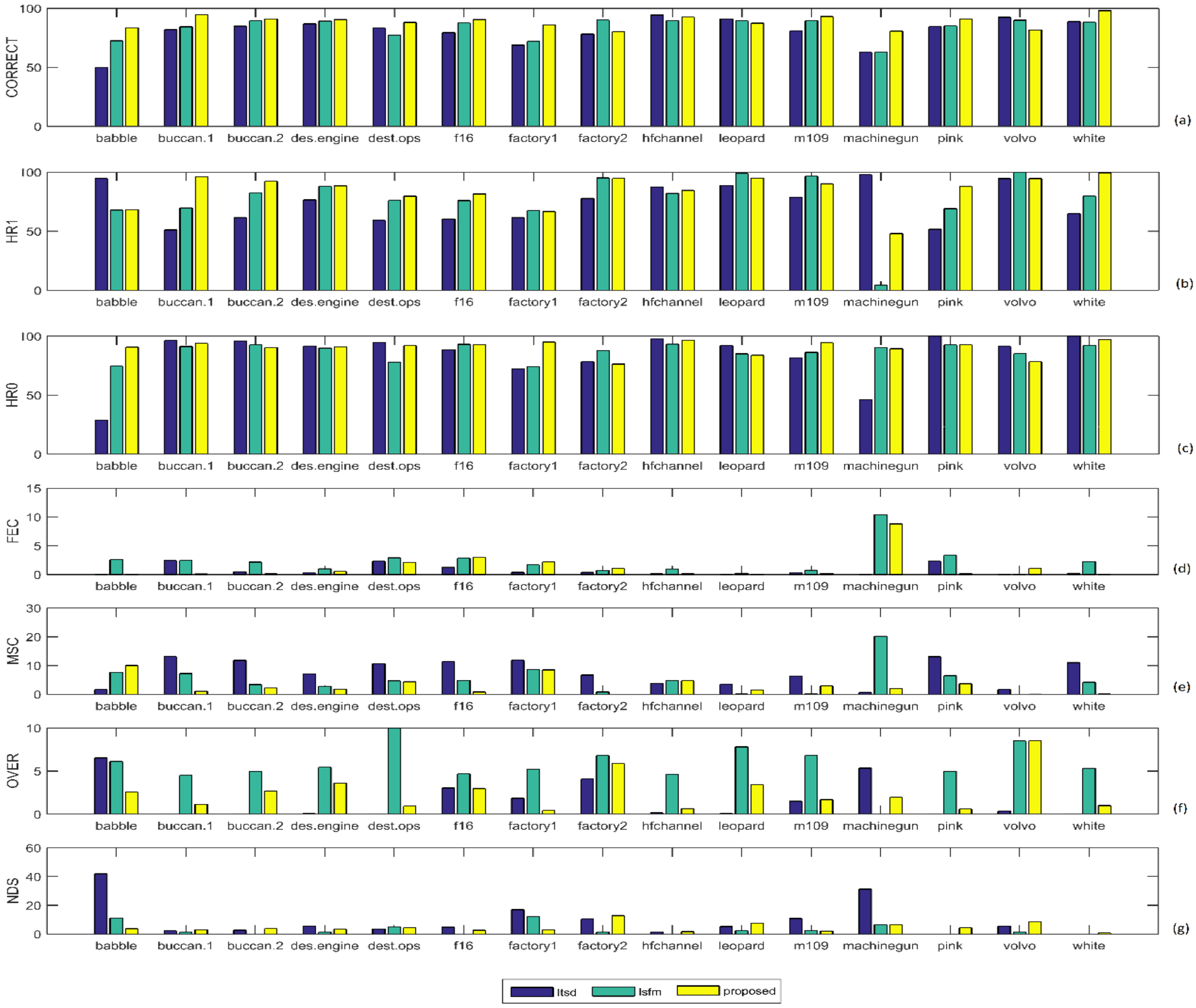

Fig.7. Accuracy and error rate comparisons of three VAD algorithms averaged over five SNR levels for 15 noises. Accuracy rate: a) CORRECT, b) HR1 and c) HR0; error rate: d) FEC, e) MSC, f) OVER and g) NDS.

CORRECT, (b) HR1 and (c) HR0. It can be seen that LTSD performs lower than the others in all three parameters CORRECT, HR1 and HR0. It performs similar to LSFM in HR1 but, suffers degradation of HR0 with increase in SNR level. LSFM performs average in all cases. Both LTSD and LSFM show gradual increase in CORRECT and HR1 but false acceptance rate of non-speech region increases since HR0 decreases with increase in SNR level. Our proposed method performs much better than the other two and also HR0 remains almost same in all SNRs. As for the error rate, from (d), (e), (f) and (g) we can see that LTSD achieve the best performance in FEC and OVER while it suffers in MSC and NDS, because of its low HR0. LSFM suffers from false positive which can be seen in OVER and FEC. On average our proposed method perform better as compared to the other two and also it achieve the best result in MSC for all SNR levels.

Table 1 provides the average performance of the three VAD algorithms over 15 noises and 5 SNR levels. We can verify that our proposed method achieves 88.61 % CORRECT which is 4.75 and 8.05 % higher than that of LSFM and LTSD, respectively. For speech hit rate our proposed method yields 84.54 % which is 7.58 and 10.68 % higher than that of LSFM and LTSD, respectively. And for non-speech hit rate our proposed method yields 90.30 % which is 3.19 and 6.59 % higher

Table 1. Average Performance Comparison for All 15 Noises over Five SNR Levels

VAD	LTSD	LSFM	PROPOSED
CORRECT	80.56	83.86	*88.61*
HR1	73.86	76.96	*84.54*
HR0	83.71	87.11	*90.30*
FEC	*0.71*	2.29	1.32
MSC	7.65	5.08	*2.93*
OVER	*1.53*	5.71	2.53
NDS	9.54	*3.05*	4.60

Note: The italicized numbers represent the best performance among all compared algorithms.

Copyright © 2017 MECS

than that of LSFM and LTSD, respectively. And among all the four error rate parameters our proposed method attains best result only in MSC – 2.93 % which is 2.15 and 4.72 % lower than that of LSFM and LTSD respectively. LSFM achieves the best in NDS i.e. 3.05 % and the remaining two parameter by LTSD i.e. 0.71 % FEC and 1.53 % OVER.

Figure 7 provide the accuracy and error rate comparisons of three evaluated algorithms averaged over five SNR levels for all 15 noises. From fig. 7(a) it can be clearly seen that in terms of CORRECT our proposed method score more than the other two standard algorithms in 11 out of 15 noises on average by 5 %. For other four noises- factory2, hfchannel, leopard and Volvo noise, LSFM and LTSD performs better than our proposed method. This might be due to mismatch of R and M values. Machinegun noise is considered to be highly non-stationary as it contains firing and silence at irregular intervals. Even in this noise our method performs relatively well. Overall we can see that LSFM performs moderate and LTSD is the worst among the three VAD's. For speech hit rate shown in fig. 7(b) there are some noises where LTSD is better than the other two. Especially for speech babble noise, machinegun noise and high frequency channel noise using a long term information based on LTSD VAD algorithm is suitable for low SNRs. From fig. 7(c) we can see that in terms of non-speech hit rate LSFM performs almost similar to our proposed method but with a slight difference of 5% on average. For error rate from fig. 7(d) to fig. 7(g) we observed that LTSD performs the best in terms of FEC and OVER. But, it produces high false acceptance in MSC due to its noise spectrum averaging property. LSFM and proposed method gives quite high error rate in OVER as compared to LTSD. Our proposed method achieves the best in MSC and yields a moderate behavior in NDS. Low MSC is important for any application since high MSC implies that speech frames are detected as non-speech. The MSC score of our proposed method is lower than LSFM and LTSD score for 10 noises and more for babble, hfchannel, leopard, m109 and machinegun noise. LSFM obtain the best performance in NDS while LTSD yields poor results. Overall considering all the four error rate evaluation metric we can conclude that our proposed method is better than the other compared algorithms.

V. Conclusion

In this paper, a new VAD algorithm is presented based on long term dominant frequency and spectral flatness measure. The proposed algorithm is intended to improve the robustness of decision mechanism by reducing false positive suffer by most algorithms. Decision rule using both LSFM and dominant frequency components are also discussed and a new spectral envelope based on dominant frequency is introduced to maximize the discriminative power. Experiments are carried out using clean-speech test set of CSTR, University of Edinburgh and all 15 noises of NOISEX92 database at five different SNR levels (-10 dB, -5 dB, 0 dB, 5 dB, 10 dB). Performance comparison are done against two standard algorithms - LTSD and LSFM. Experimental results show that our proposed algorithm outperforms the other two algorithms in terms of accuracy rate. While for error rate LTSD is more robust however, our proposed method also achieve moderate result. Moreover our proposed method achieve the lowest MSC among the compared algorithms, since its increase will lead to miss in speech region. Further improvement can be done by fine tuning the initial parameters required for computing each features.

References

[1] J. Górriz, J. Ramírez, E. W. Lang, C. G. Puntonet, and I. Turias, "Improved likelihood ratio test based voice activity detector applied to speech recognition," *Speech Communication*, vol. 52, no. 7, pp. 664–677, 2010.

[2] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[4] D. Freeman, G. Cosier, C. Southcott, and I. Boyd, "The voice activity detector for the pan-european digital cellular mobile telephone service," pp. 369–372, 1989.

[5] D. Enqing, Z. Heming, and L. Yongli, "Low bit and variable rate speech coding using local cosine transform," vol. 1, pp. 423–426, 2002.

[6] J. Alam, P. Kenny, P. Ouellet, T. Stafylakis, and P. Dumouchel, "Supervised/unsupervised voice activity detectors for text-dependent speaker recognition on the rsr2015 corpus," 2014.

[7] Benyassine, E. Shlomot, H.-Y. Su, and E. Yuen, "A robust low complexity voice activity detection algorithm for speech communication systems," pp. 97–98, 1997.

[8] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Labs Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.

[9] T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," *Entropy*, vol. 2, no. 2.5, p. 3, 2005.

[10] D.-J. Liu and C.-T. Lin, "Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, pp. 609–621, 2001.

[11] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical v/uv classification algorithm," *IEEE Transactions on speech and audio processing*, vol. 7, no. 3, pp. 333–338, 2010.

[12] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3, pp. 271–287, 2004.

[13] T. Fukuda, O. Ichikawa, and M. Nishimura, "Long-term spectro-temporal and static harmonic features for voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 834–844, 2010.

[14] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.

[15] Y. Ma and A. Nishihara, “Efficient voice activity detection algorithm using long-term spectral flatness measure,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, p. 87, 2013.

[16] T. V. Pham, C. T. Tang, and M. Stadtschnitzer, “Using artificial neural network for robust voice activity detection under adverse conditions,” pp. 1–8, 2009.

[17] P. Estevez, N. Becerra-Yoma, N. Boric, and J. Ramırez, “Genetic programming-based voice activity detection,” *Electronics Letters*, vol. 41, no. 20, pp. 1141–1143, 2005.

[18] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, “Applying support vector machines to voice activity detection,” vol. 2, pp. 1124–1127, 2002.

[19] G. A. N. Anita Ahmad, Fernando Soares Schlindwein, “Comparison of computation time for estimation of dominant frequency of atrial electrograms: Fast fourier transform, blackman tukey, autoregressive and multiple signal classification,” *Biomedical Science and Engineering*, pp. 843–847, 2010.

[20] C. Valentini-Botinhao *et al.*, “Superseded-noisy speech database for training speech enhancement algorithms and tts models,” 2016.

[21] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113 – 120, 1979.

[22] A. Davis, S. Nordholm, and R. Togneri, “Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 412–424, 2006.

[23] K. Sjölander and J. Beskow, “Wavesurfer-an open source speech tool.” pp. 464–467, 2000.

[24] A. Varga and H. J. Steeneken, “Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech communication*, vol. 12 , no. 3, pp. 247–251, 1993.

Authors’ Profiles

Naorem karline Singh has completed B-tech in computer science and engineering, in 2015 from NIT Manipur, India and currently pursing his M-tech degree at the same institute.

His main area of interests are speech processing, computer networking and web development.

Yambem Jina Chanu has received Ph.D degree in computer science and engineering from NERIST, Itanagar, India. She is currently working as an assistant professor in NIT Manipur.

Her main area of interests are image processing, steganography, pattern recognition, image segmentation and speech processing.

How to cite this paper: Naorem Karline Singh, Yambem Jina Chanu,"Robust Voice Activity Detection Algorithm based on Long Term Dominant Frequency and Spectral Flatness Measure", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.9, No.8, pp.50-58, 2017.DOI: 10.5815/ijigsp.2017.08.06