Interspeech 2018
2-6 September 2018, Hyderabad

Robust Voice Activity Detection Using Frequency Domain Long-Term Differential Entropy

Debayan Ghosh[1]*, R. Muralishankar*[2]*, and Sanjeev Gurugopinath*[1]

[1]Dept. of ECE, PES University, Bengaluru 560085, India
[2]Dept. of ECE, CMR Institute of Technology, Bengaluru 560037, India
gdebayan95@gmail.com, muralishankar@cmrit.ac.in, sanjeevg@pes.edu

Abstract

We propose a novel voice activity detection (VAD) scheme employing differential entropy at each frequency bin of power spectral estimates of past and present overlapping speech frames. Here, the power spectral estimate is obtained by employing the Bartlett-Welch method. Later, we add entropies across frequency bins and denote this as the frequency domain long-term differential entropy (FLDE). Long-term averaging enhances VAD performance under low signal-to-noise-ratio (SNR). We evaluate the performance of proposed FLDE scheme, considering 12 types of noises and 5 different SNRs which are artificially added to speech samples from the SWITCHBOARD corpus. We present VAD performance of FLDE and compare with existing VAD algorithms, such as ITU-T G.729B, likelihood ratio test, long-term signal variability, and long-term spectral flatness measure based algorithms. Finally, we demonstrate that our FLDE-based VAD performs with best average accuracy and speech hit-rate among the VAD algorithms considered for evaluation.

1. Introduction

An important problem in many speech processing applications, is to identify speech or non-speech regions in a given audio signal. Signals from non-speech regions are mostly silence, noisy and/or speech like (in the case of babble), which cannot be comprehended. Determination of speech or non-speech intervals poses many challenges due to the non-stationary nature of speech and/or background noises. An algorithm employed to detect the presence or absence of speech is referred to as voice activity detection (VAD). VAD is crucial in tasks such as automatic speech recognition (ASR) [1] and speaker diarization [2, 3]. VAD is also used in voice over internet protocol (VoIP) applications [4, 5], where only segments determined to be speech are transmitted over the channel, thereby saving bandwidth by avoiding unnecessary transmissions. In most VAD algorithms, performance trade-offs are studied by maximizing detection of speech intervals while minimizing the false detection of non-speech regions as speech.

The core of any VAD algorithm consists of two parts: *feature extraction* and a *decision mechanism*. In the first part, we extract information from a given speech signal through its parameters, which has discriminative characteristics between speech and non-speech. Using these features and a set of decision rules, speech/non-speech decisions are made in the second part. A variety of features have been proposed for VAD. Initially, frame wise short-term energy [6] and number of zero crossings [6] as a feature were employed. However, the performance of these algorithms deteriorate under low signal-to-noise-ratio (SNR) and non-stationary background noise. More advanced features such as order statistics [7], likelihood ratio test (LRT) [8], energy in subbands ($0-1$ kHz, $1-2$ kHz, $2-3$ kHz, $3-4$ kHz) [6] and autocorrelation [9] were proposed, however, these algorithms assume the background noise process to be stationary. In all these algorithms, the signal is divided into non-overlapping frames of 20 ms, and VAD is performed on each frame independently. Most of the above algorithms work well under high SNR ($>$ 10 dB), but offer a poor performance in the presence of non-stationary noises.

We can discriminate speech and non-speech sounds based on their different variability profiles. It is known from the literature that a person with an average speaking rate produces approximately 10 to 15 phonemes per second [10]. Each of these phonemes presents different spectral distribution, and therefore, the statistics of speech vary significantly over time. This necessitates a mechanism to analyze signal over long duration differing with usual short-time analysis. The use of long windows was exploited in long-term signal variability (LTSV) [11], and in long-term spectral flatness measure (LSFM) [12] algorithms. These algorithms show a drastic improvement under low SNR ($<$ 10 dB) and in the presence of non-stationary noises, as opposed to their short-term counterparts.

In this work, we propose a frequency domain long-term differential entropy (FLDE) of speech signal as a feature to discriminate between speech/non-speech regions under low SNR and non-stationary noise conditions. The differential entropy (DE) was successfully used to find the presence or absence of a primary user activity in spectrum sensing for cognitive radios [13]. This motivates us to employ DE with long-term analysis similar to LSFM [12] and LTSV [11]. Here, we divide the signal into overlapping frames, with window length of 20 ms and an overlap of 10 ms. We then average the spectral estimates employing the Bartlett-Welch method [11, 12] from M consecutive frames. We assume that samples from each frequency bin are Gaussian distributed, and compute the sample variance over each frequency bin to find an estimate of the DE. The computed DE values are summed across frequency bins to get the FLDE feature. Our approach in extraction of FLDE feature differs from LTSV as follows: (i) sample variance of the spectral estimates to compute differential entropy over each frequency bin (ii) summing the entropy estimates across frequency bins. We experimentally evaluate and compare the performance of the FLDE-based VAD in presence of (i) different noise types, and (ii) SNR conditions, over LTSV, LSFM, LRT and G.729B [14] algorithms. The proposed FLDE-based VAD outperforms in the comparative analysis with improved correct speech detection rate, while maintaining a similar – if not worse – false-alarm rate. The remainder of the paper is organized as follows. Differential entropy as a measure is introduced in § 2. The FLDE-based VAD algorithm is proposed in § 3. Experimental results are discussed in § 4, and conclusions are drawn in § 5.

10.21437/Interspeech.2018-1431

2. The Differential Entropy Measure

The differential entropy $h(\mathcal{X})$ of a Gaussian random variable $\mathcal{X}$, with mean μ, variance σ^2 and PDF $f_{\mathcal{X}}(x)$ is given by

$$h(\mathcal{X}) \triangleq -\int_{-\infty}^{\infty} f_{\mathcal{X}}(x)\log(f_{\mathcal{X}}(x))dx = \frac{1}{2}\log(2\pi e\sigma^2).$$

In this work, we estimate the DE in a given set of M observations $\{X_1,\ldots,X_M\}$, assumed to be independent and identically distributed samples from the Gaussian. The sample mean and variance are respectively given by $\widehat{X} = \frac{1}{M}\sum_{i=1}^{M} X_i$, and $\widehat{\sigma^2} = \frac{1}{M-1}\sum_{i=1}^{M}(X_i - \widehat{X})^2$. The required maximum likelihood estimate (MLE) of DE in $\{X_1,\ldots,X_M\}$ is given by [13]

$$\widehat{h}(\mathcal{X}) = \frac{1}{2}\log\left\{\frac{2\pi e}{M-1}\sum_{i=1}^{M}(X_i - \hat{X})^2\right\}. \quad (1)$$

3. Frequency Domain Long-Term Differential Entropy (FLDE)-based VAD

In this section, we describe steps involved in extraction of proposed FLDE feature from a given speech/non-speech signal, $x(n), n = 1,\ldots,N_w$, where, N_w is frame length. We find the short-term Fourier transform (STFT) of $x(n)$ over N_w samples, and with frame shift N_{sh} to obtain

$$X(m,\omega_k) = \sum_{l=(m-1)N_{sh}+1}^{N_w+(m-1)N_{sh}} w(l-(m-1)N_{sh}-1)x(l)e^{-j\omega_k l}, \quad (2)$$

where $X(m,\omega_k)$ is the STFT coefficient at frequency $\omega_k, k = 0,1,\ldots,(N_{DFT}/2)-1$, of the m^{th} frame, $m = 1,2,\ldots$. Also, N_{DFT} is the length of discrete Fourier transform (DFT) sequence used in the spectral estimate of $x(n)$, and $w(l), l = 1,\ldots,N_w$ represents the Hann window of length N_w. Then, we estimate the short-time power spectrum $S(n,\omega_k)$ using the Bartlett-Welch method [12], where the squared magnitudes of STFT under M consecutive frames are averaged as:

$$S(n,\omega_k) = \frac{1}{M}\sum_{m=n-M+1}^{n}|X(m,\omega_k)|^2, \quad (3)$$

with a new frame index $n = 1,2,\ldots$. In the next step, we find the variance of power spectrum $S(n,\omega_k)$ at each frequency bins, ω_k, for the past R frames with respect to a current frame of interest. This is done by finding sample mean, $\widehat{S}(p,\omega_k) = \frac{1}{R}\sum_{n=p-R+1}^{p} S(n,\omega_k)$, the sample variance, $\widehat{\sigma^2}(p,\omega_k) = \sum_{n=p-R+1}^{p}(S(n,\omega_k) - \widehat{S}(p,\omega_k))^2$, and MLE of DE:

$$\widehat{h}(p,\omega_k) = \frac{1}{2}\log\left\{\frac{2\pi e}{R-1}\widehat{\sigma^2}(p,\omega_k)\right\}. \quad (4)$$

Finally, the FLDE feature, $L_x(p)$, for the p^{th} indexed frame is extracted by summing of DE across frequency bins, ω_k, as

$$L_x(p) = \sum_k \widehat{h}(p,\omega_k). \quad (5)$$

The proposed algorithm decides speech in the p^{th} frame, when $L_x(p) > \tau_p$, and decides non-speech, otherwise, where τ_p is the detection threshold; which will be discussed later.

Table 1: *Optimum (M,R) values for FLDE-based VAD*

Noise	- 10 dB	-5 dB	0 dB	5 dB	10 dB
Speech Babble	(30,30)	(30,30)	(10,30)	(10,30)	(1,30)
White	**(5,30)**	**(5,30)**	(5,20)	(5,10)	(1,10)
Pink	(10,30)	**(5,30)**	(5,20)	(10,10)	(5,10)
Jet Cockpit	(1,30)	(1,30)	**(5,30)**	(5,10)	(10,10)
F-16 Cockpit	(20,30)	(1,30)	(10,10)	(1,5)	(5,5)
Factory Floor	(30,10)	(30,10)	(1,30)	(20,10)	(10,20)
HF Channel	**(5,30)**	(1,30)	(5,10)	(10,10)	(5,20)
Military Tank	**(5,30)**	**(5,30)**	(10,10)	(5,5)	(5,5)
Military Vehicle	(10,10)	(10,10)	(10,5)	(5,10)	(1,5)
Vehicle Interior	(5,5)	(5,5)	(1,10)	(1,5)	(1,5)
Machine Gun	(1,5)	(1,5)	(20,5)	(1,5)	(1,20)
Destroyer Engine	**(5,30)**	**(5,30)**	(5,20)	(1,20)	(5,10)

Parameters of the FLDE-based VAD

The FLDE is extracted by appropriately setting parameter values as follows. The N_w, and N_{sh} are set to 20 ms and 10 ms, respectively. The sampling rate, $f_s = 8$ kHz. Input signal is Hann windowed, followed by an $N_{DFT} = 512$ point DFT. The frequency range is set from 500 Hz to 4 kHz [11, 12]. The start and end frequency bins, k_s and k_e, are evaluated using the relation, $k_s = N_{DFT}(\frac{500}{fs})$ and $k_e = N_{DFT}(\frac{4000}{fs})$, respectively.

Further, parameters M and R are chosen as follows. The total misclassification error employing the SWITCHBOARD corpus [15] was computed for all combinations of $M(1,5,10,20,30)$, and $R(5,10,20,30)$, with 12 different noises and five SNR conditions. Optimal choice of the pair (M,R) is highlighted in Table 1, from where we choose the combination of $(5,30)$, which appears most frequently. A similar approach was used to find the best (M,R) for the LSFM and LTSV algorithms. The choice of $(10,30)$ for the LSFM, and $(20,30)$ for LTSV were determined using the TIMIT database [16]. For a fair comparison, we checked the best (M,R) for LSFM and LTSV on the SWITCHBOARD corpus, which yield the same (M,R) pairs obtained with TIMIT database. Therefore, the choice of (M,R) depends more on the background noise than the database under consideration.

The efficacy of a VAD algorithm depends on the choice of the detection threshold. For example, in Fig. 1, we can observe variation in FLDE feature under speech and non-speech regions. This necessitates an adaptive threshold – as opposed to a fixed threshold – under varying acoustic conditions. For a given signal, the FLDE feature is computed after initial $M + R - 1$ frames. For e.g., if $M = 5$, $R = 30$, the first feature is calculated after 34 frames, or 0.34 s. We assume that the first 1.34 s of input has only non-speech signal, and 100 realizations of the FLDE feature in this time is computed and saved in a buffer, Ψ_{INL}. An estimate of the initial threshold can be evaluated as $\tau_{INL} = k\min(\Psi_{INL})$, where, $k \in (0.75,1]$, is a scaling factor. Then, we maintain two buffers, namely, Ψ_N, which contains the FLDE values of the last 100 frames from the decided non-speech regions, and Ψ_{S+N}, which contains the FLDE values of the last 100 frames from the decided speech plus noise regions. The threshold, τ_m, for the m^{th} frame is computed as follows.

$$\tau_m = \alpha\min(\Psi_{S+N}) + (1-\alpha)\max(\Psi_N) \quad (6)$$

where $\alpha \in (0,1)$ is a tuning parameter. Experimentally, it was found that $\alpha = 0.45$ resulted in least misclassification error under the SWITCHBOARD corpus.

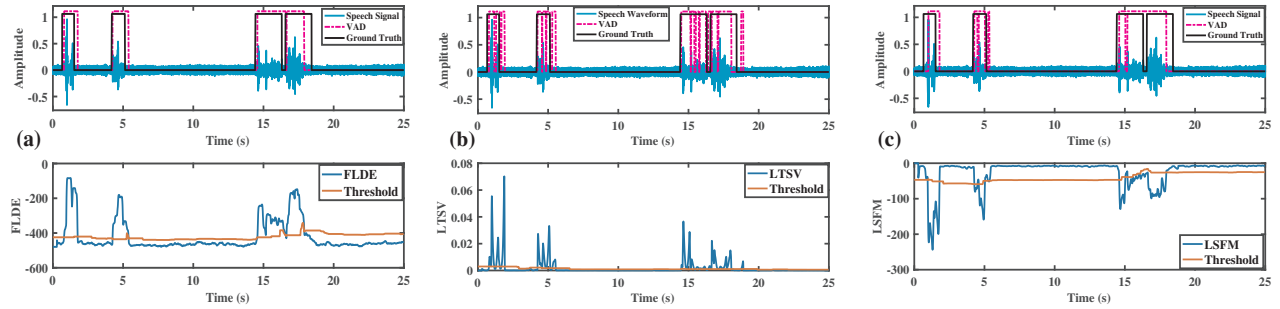

Figure 1: *VAD decisions at* 0 *dB under Destroyer engine noise background: (a) FLDE (b) LTSV (c) LSFM.*

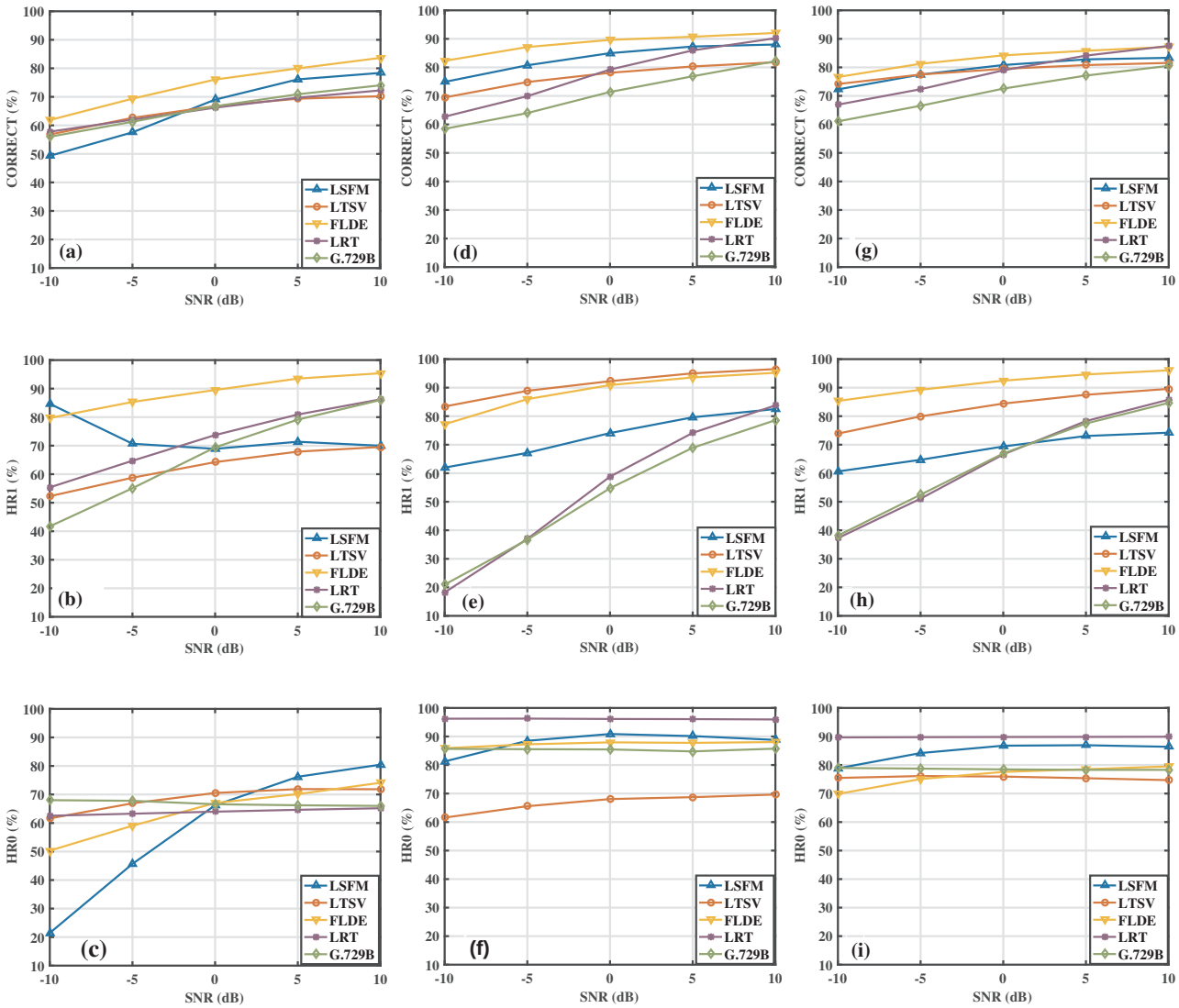

Figure 2: *Performance of LSFM, LTSV, FLDE, LRT and G.729B using the evaluation parameters* $CORRECT(\%)$*, Speech hit-rate,* $HR1(\%)$ *and Non-speech hit-rate* $HR0(\%)$*, with five SNRs under the following noises: (a-c) Babble and (d-f) Destroyer engine. Average of performances (g)* $CORRECT(\%)$ *(h)* $HR1(\%)$ *(i)* $HR0(\%)$ *across* 12 *noise backgrounds.*

4. Results and Discussion

4.1. Performance Evaluation of FLDE-based VAD

We evaluate performance of the proposed FLDE-based VAD algorithm and compare its performance with four reference VAD algorithms using the SWITCHBOARD corpus, under 12 different noises from the NOISEX-92 database [17], and at 5 different SNRs, $-10, -5, 0, 5$ and 10 dB. This corpus is composed of approximately 2400 telephone conversations between unacquainted adults, from which we used 46 conversations that were

syntactically parsed. Each conversation is of 1 minute duration, resulting in a total evaluation period of 46 minutes. The reference labels for this corpus are computed using the available manually segmented labels corresponding to each of 46 files, with a decision every 10 ms. The sampling frequency employed in this corpus is 8 kHz. Further, this corpus consists of both long and short bursts of silence and speech, representing a more realistic scenario to evaluate different VAD algorithms. One may observe that the TIMIT database [16] was employed in [11] and [12] where test samples were prepared by introducing pauses with known duration between different sample utterances from TIMIT. To account for unpredictable nature in a conversation, we choose our test samples from SWITCHBOARD corpus over TIMIT. The 12 noise samples from the NOISEX-92 database

Table 2: *Average of Speech hit-rate, $HR1$ and Non-speech hit-rate, $HR0$ under different background noises and five SNRs.*

Performance under White, Pink, Car, HF-channel and Tank noises					
Algorithm	-10 dB	-5 dB	0 dB	5 dB	10 dB
LSFM	60.2, 86.3	66.9, 89.6	72.6, 91.0	76.9, 89.8	78.7, 88.7
LTSV	75.8, 83.5	82.6, 83.3	87.8, 82.0	91.1, 80.8	93.8, 79.8
FLDE	**83.1**, 80.0	**87.5**, 85.6	**91.5**, 86.6	**93.6**, 87.1	**95.4**, 87.1
LRT	29.8, **97.5**	44.3, **97.4**	61.8, **97.4**	75.6, **97.3**	84.3, **97.2**
G.729B	28.3, 90.3	45.1, 90.1	61.8, 89.9	74.0, 89.7	82.4, 89.6
Performance under Babble, Jet-cockpit, F-16 and Factory floor noises					
Algorithm	-10 dB	-5 dB	0 dB	5 dB	10 dB
LSFM	62.7, 70.2	65.1, 79.4	69.9, 84.8	74.2, 86.9	75.3, 87.0
LTSV	68.6, 74.8	75.8, 76.4	81.4, 77.1	85.6, 76.8	87.7, 76.0
FLDE	**84.4**, 61.5	**88.2**, 69.4	**91.2**, 75.0	**94.4**, 76.7	**95.7**, 79.5
LRT	29.8, **87.2**	45.6, **87.4**	62.7, **87.6**	75.6, **87.7**	84.1, **87.8**
G.729B	28.8, 85.7	45.8, 85.6	62.8, 85.0	74.8, 84.6	83.0, 84.4
Performance under Destroyer engine, Machine-gun and Military vehicle noises					
Algorithm	-10 dB	-5 dB	0 dB	5 dB	10 dB
LSFM	58.7, 77.9	60.7, **81.6**	63.4, **82.4**	65.1, **82.2**	65.4, **81.7**
LTSV	77.8, 62.9	80.9, 63.8	82.9, 64.4	84.1, 64.5	85.1, 64.5
FLDE	**90.3**, 64.1	**93.7**, 65.1	**95.7**, 66.0	**96.8**, 66.6	**97.6**, 66.8
LRT	59.6, **79.9**	69.8, 80.0	79.6, 80.1	86.3, 80.3	90.6, 80.5
G.729B	67.1, 51.0	73.8, 50.7	81.0, 50.7	86.7, 50.8	90.7, 51.3

are broadly classified into (i) stationary (White, Pink, Car, HF-channel and Tank) (ii) non-stationary (Babble, Jet-cockpit, F-16 and Factory floor) and (iii) non-stationary and heavy-tailed (Destroyer engine, Machine-gun and Military vehicle) noises. Figure 1 shows VAD decisions, feature and a running threshold of long-term based algorithms, (a) proposed FLDE, (b) LTSV, and (c) LSFM. Note that decisions from FLDE are smoother and closer to the *Ground Truth*, when compared with LTSV and LSFM. As done in [12], we evaluate the performance of our VAD algorithm based on the three metrics:

1. Correct Detection ($CORRECT$): Regions correctly classified as speech and non-speech. $CORRECT = (N_{1,1} + N_{0,0}) / \left(N_1^{\text{ref}} + N_0^{\text{ref}}\right)$.
2. Speech hit-rate ($HR1$): Regions correctly detected as speech, in all speech regions. $HR1 = N_{1,1}/N_1^{\text{ref}}$.
3. Non-speech hit-rate ($HR0$): Regions correctly detected as non-speech, in all non-speech regions. $HR0 = \frac{N_{0,0}}{N_0^{\text{ref}}}$,

where N_1^{ref} and N_0^{ref} are the numbers of speech and non-speech regions in the database, respectively, while $N_{1,1}$ and $N_{0,0}$ are the numbers of speech and non-speech regions which are correctly classified. Ideally, we expect a VAD algorithm to perform well in all three counts. However, most algorithms maximize $HR1$, with a minimum false-alarm ($1 - HR0$).

4.2. Comparative Performance of VAD algorithms

The VAD decisions are computed for every 10 ms interval and tagged as speech or non-speech frame. If a frame has both speech and non-speech information, then, it will be declared as speech. By this, we may avoid discarding speech at the beginning and at the end of speech plus noise regions. In Fig. 2, we present performance of the FLDE and other algorithms under Babble and Destroyer engine noise conditions with SNRs varied in steps of 5 dB from -10 to 10 dB. We make the following observations for Babble noise from Figs. 2(a), 2(b), and 2(c). First, the values of $CORRECT$ and $HR1$ is more in FLDE as compared to the other VAD algorithms, except at -10 dB where LSFM has about 4.97 % more $HR1$ than the FLDE. Next, the non-speech hit-rate, $HR0$, as shown in Fig. 2(c), are in the range of 50 to 70 % except LSFM, which has higher $HR0$ at 5 dB and 10 dB, however, it has lower $HR1$ at SNRs 5 and 10 dB. Next, Figs. 2(d), 2(e), and 2(f) present performance of the algorithms under the Destroyer engine noise. In this case, $CORRECT$ is more in FLDE than other algorithms across all SNRs. Also, FLDE is close second to LTSV in terms of $HR1$, and has a better $HR0$ than LTSV, which results in an improved overall speech/non-speech classification when compared with other VAD algorithms. Finally, an average performance across all 12 noises are shown in Figs. 2(g), 2(h) and 2(i). We can observe that FLDE outperforms all the other algorithms across all SNRs, in terms of $HR1$. However, FLDE performs marginally well in terms of $CORRECT$, except at 10 dB where the LRT scores over FLDE by about 0.31 %. In Table 2, we present the average $HR1$ and $HR0$ values for the three different classes of noise signals discussed in § 4.1. It can be observed from Table 2 that FLDE has highest average $HR1$, but with a slightly lower average $HR0$ than LRT and LSFM.

5. Concluding Remarks

In this paper, we introduced a novel VAD algorithm, namely, a frequency domain long-term differential entropy-based algorithm, where the DE was applied in a long-term sense on the DFT of the input signal. We presented the steps involved in the extraction of FLDE feature from a given speech signal. The choice of best long-term parameters, M and R, were found empirically by minimizing misclassification, $(1 - CORRECT)$. Through an extensive performance evaluation, we found that the FLDE-based VAD outperforms existing VAD algorithms, LSFM, LTSV, LRT and G.729B, especially at SNRs below 0 dB. The speech hit-rate, $HR1$ achieved by the FLDE is higher than existing algorithms with moderate false-alarm rate, $(1 - HR0)$. Our results reiterate that long-term algorithms are more robust under low SNR conditions. As a part of the future work, to ensure that the FLDE has moderate non-speech hit-rate – which needs to be maximized keeping speech-hit rate at maximum, we observe that the FLDE feature is the sum of entropies from each bin with equal weights. Herein, each bin may be weighted based on the presence or absence of speech information. This could facilitate to achieve the best possible speech hit-rate with an acceptable false-alarm rate $(1 - HR0)$.

6. References

[1] D. Vlaj, B. Kotnik, B. Horvat, and Z. Kai, "A computationally efficient mel-filter bank VAD algorithm for distributed speech recognition systems," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 4, pp. 487 – 497, March 2005.

[2] X. A. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356 – 370, February 2012.

[3] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1557 – 1565, September 2006.

[4] W. B. D. Forfang, E. Gonzalez, S. Mcclellan, and V. Viswanathan, "A VAD/VOX algorithm for amateur radio applications," in *International Journal on Advances in Telecommunications*, no. 1 and 2, 2014.

[5] Y.-C. Chang, K.-T. Chen, C.-C. Wu, and C.-L. Lei, "Inferring speech activities from encrypted skype traffic," in *Proceedings of IEEE Globecom 2008*, 2008.

[6] R. V. Prasad, A. Sangwan, H. Jamadagni, M. Chiranth, R. Sah, and V. Gaurav, "Comparison of voice activity detection algorithms for VoIP," *Proceedings ISCC 2002 Seventh International Symposium on Computers and Communications*, July 2002.

[7] R. Muralishankar, R. V. Prasad, S. Vijay, and H. N. Shankar, "Order statistics for voice activity detection in VoIP," *2010 IEEE International Conference on Communications*, March 2010.

[8] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1 – 3, January 1999.

[9] Z. Shuyin, G. Ying, and W. Buhong, "Auto-correlation property of speech and its application in voice activity detection," *2009 First International Workshop on Education Technology and Computer Science*, March 2009.

[10] A. M. Liberman, *Speech: a special code*. MIT Press, 1996.

[11] 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, March 2011.

[12] 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, July 2013.

[13] S. Gurugopinath, R. Muralishankar, and H. N. Shankar, "Spectrum sensing for cognitive radios through differential entropy," *EAI Endorsed Transactions on Cognitive Communications*, vol. 2, no. 6, May 2016.

[14] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit, "ITU-T recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 64 – 73, 1997.

[15] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ser. ICASSP'92. Washington, DC, USA: IEEE Computer Society, 1992, pp. 517–520. [Online]. Available: http://dl.acm.org/citation.cfm?id=1895550.1895693

[16] J. Garofolo, N. I. of Standards, and T. (U.S.), *Darpa Timit: Acoustic-phonetic Continuous Speech Corps CD-ROM*. U.S. Department of Commerce, National Institute of Standards and Technology, 1993. [Online]. Available: https://books.google.co.in/books?id=2VCAHAAACAAJ

[17] A. Varga and H. J. M. 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 Commun.*, vol. 12, no. 3, Jul. 1993.