A COMPARISON OF FEATURES FOR VOICE ACTIVITY DETECTION - A REVIEW AND SOME EXPERIMENTAL RESULTS

J. Honnatteppanavar[1] and B. G. Nagaraja
[1]Visvesvaraya Technological University, Belagavi.
[2]Dept. of Electronics & Communication Engg., Jain Institute of Technology, Davangere, Karnataka

ABSTRACT

Speech/voice activity detection (VAD) is an essential front end step for numerous speech processing applications to classify voiced and unvoiced regions. Over the years, many algorithms/features have been proposed for the VAD. Most of these algorithms provide unsatisfactory performance due to the presence of background noise. An ideal VAD needs to be independent of noise condition and application area. Thus, choosing a suitable feature/method for obtaining appropriate decision for VAD is an essential task. This paper reports a concise experimental review of four methods for VAD under background noise conditions. The different VAD studies are carried out using noisy speech corpus (NOIZEUS). Comparative results indicate that the spectral energy based VAD with adaptive threshold are considerably more useful in low signal-to-noise-ratio (SNR) condition.

Keywords: *VAD, accuracy, ZCR, Spectral energy, SNR.*

Introduction

Voice activity detection (VAD) is a binary classification problem of separating speech/voiced segments from background noisy speech [1]. In digital cellular telecommunication systems, VAD is used to detect unvoiced frames and as a result average bit rates are reduced [2]. Further, VAD may also enhance the accuracy of automatic speech recognition system by correctly identifying the voiced boundaries in a speech signal [3]. A general VAD technique includes mainly two important steps: feature extraction and a classifier. Literature reveals that many VAD algorithms have been proposed. The major diversity among most of the state-of-the-art methods is the features considered [4]. Figure 1 shows the general block diagram of the VAD method.

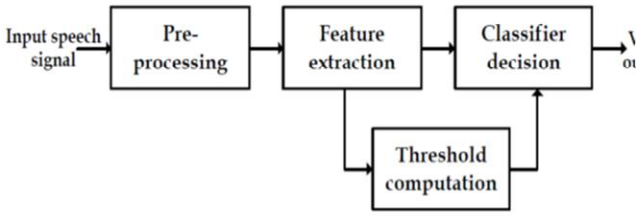


Figure 1: Block diagram of the VAD method.

The work in [5], proposed a robust VAD to classify speech and non-speech in different noisy conditions based on power spectral deviation. To better represent the power spectral deviation, teager energy features were used. The proposed VAD was evaluated under various noise condition (babble, office and car noise) and signal to noise ratio (SNR) levels, viz., 0, 5, 10, 15 dB. The performance evaluation using total error rate, false acceptance rate and false rejection rate showed that the proposed VAD achieved better accuracy than the traditional VAD algorithms. The short-time spectrum energy in the overlapped speech frames based VAD was discussed in [6]. The noise signal energy from the higher frequency band (2.5 kHz to 4 kHz) was subtracted from the lower frequency band noisy speech spectrum and then to smooth the speech spectrum, a moving average filter was used. It was observed that the proposed VAD achieved better accuracy for negative SNR levels (-5 dB and -10 dB).

In the rest of this paper, Section 2 describes the different methods for VAD. A database used for the study and the experimental setups are provided in Section 3. Section 4 discusses the VAD results for different types of noise and SNR levels. The paper concludes in Section 5.

Features for VAD

Zero crossing rate (ZCR): In discrete-time-signal, a zero crossing occur if successive samples have distinct algebraic signs (`+' and `-'). The rate at which crossings occur is the ZCR of the speech frame. Figure 2 depicts the ZCR. For a given speech frame (x_n) of M samples, ZCR is defined as [7]:

$$ZCR_n = \frac{1}{2M}\sum_{i=1}^{M}[|sgn(x_n(i)) - sgn(x_n(i-1))|]$$

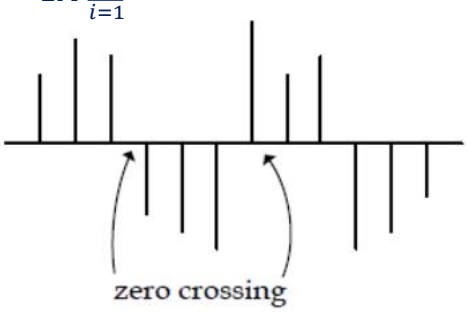


Figure 2: ZCR in discrete-time-signals.

Short-time energy (STE): The short-time energy of the frame,x_n is defined as

$$STE_n = \sum_{i=1}^{M}(x_n(i)h(n-i)^2$$

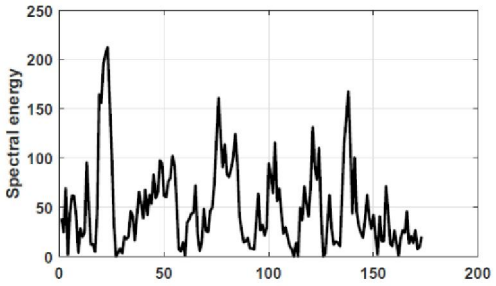
where, $h(n)$ is the hamming window.

Spectral entropy: Primary feature for VAD in the frequency domain is the spectral entropy and is defined as [8]:

$$SE(l) = -\sum_{i=1}^{M}\tilde{\phi}_{xx}(i,l) \times \log\left(\tilde{\phi}_{xx}(i,l)\right)$$

where, $\tilde{\phi}_{xx}(i,l)$ is the normalized spectrum. The SE reflects the flatness of the speech signal spectrum.

Spectrum energy: Figure 3 shows the spectral energy of the noisy speech signal from the NOIZEUS database at 0 dB and 5 dB SNR levels [9, 10]. It is necessary that for a low SNR noisy speech signal, an adaptive threshold value is necessary to classify the voiced and unvoiced regions. Figure 4 shows the block diagram of the adaptive spectral energy based VAD method.



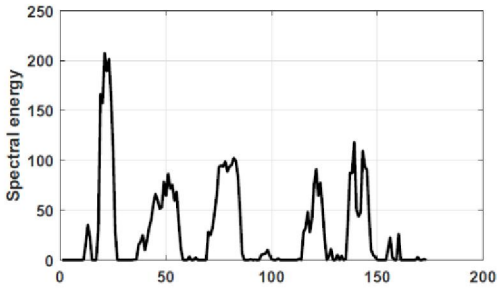


Figure 3: Spectral energy diagram for noisy speech at 0 dB and 5 dB SNR (a) 0 db SNR (b) 5 db SNR

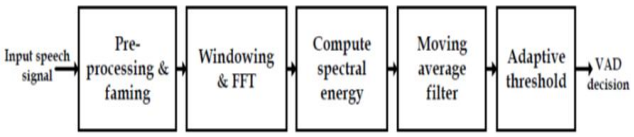


Figure 4: Block diagram of the adaptive spectral energy based VAD method.

Experimental set-up

NOIZEUS noisy speech corpus [9, 10] was used in the experiments. We chose one male speaker speech data ("*He knew the skill of the great young actress*") for the evaluation of four different methods. Three different noise conditions, viz., airport, babble and train at 0 dB and 5 dB SNR levels are considered for the study. Speech signals are sampled at 8 kHz with frame duration of 30 ms (50% overlap) is used. To assess the classification accuracy of the VADs, three metrics (unvoiced, voiced and average hit rates) are used [11].

UHR =No. of unvoiced frames correctly classified to the no. of actual unvoiced frames.

VHR =No. of voiced frames correctly classified to the no. of actual voiced frames.

$$\text{AHR} = \frac{1}{2}(\text{UHR} + \text{VHR}).$$

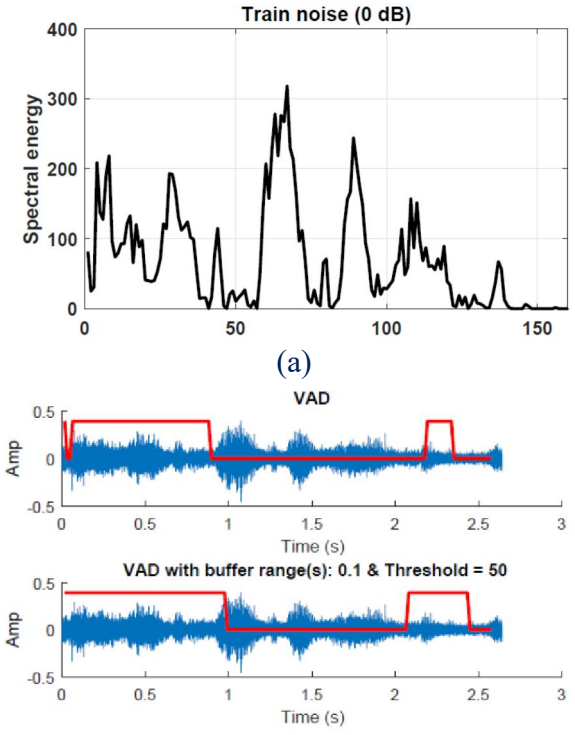
VAD results

Table 1 shows the performance of four VAD methods for different types of noise and SNR levels. Some of the observations are as follows: The performance of ZCR-VAD is comparable to that of STE-VAD. However, when the noise level increases, then few of the genuine voiced frames are missed in the classification. In most of the cases, the SEN-VAD achieved the better performance when compared with other three methods. This may be due to the use of

adaptive threshold value based on the spectral energy (Figure 5).

Table 1: Experimental results for different types of noise and SNR levels.

Noise type	ZCR-VAD	STE-VAD	SE-VAD	SEN-VAD
Airport (5 dB)	UHR=51.40 VHR=89.12 AHR=70.26	UHR=50.40 VHR=90.20 AHR=70.30	UHR=52.40 VHR=91.42 AHR=71.91	UHR=53.4 VHR=92.80 AHR=73.10
Babble (5 dB)	UHR=51.23 VHR=90.40 AHR=70.81	UHR=51.40 VHR=91.20 AHR=71.30	UHR=53.20 VHR=92.01 AHR=72.60	UHR=53.82 VHR=92.94 AHR=73.38
Train (5 dB)	UHR=50.40 VHR=89.45 AHR=69.92	UHR=51.20 VHR=90.40 AHR=70.80	UHR=52.80 VHR=91.40 AHR=72.10	UHR=53.42 VHR=92.56 AHR=72.99
Airport (0 dB)	UHR=46.80 VHR=87.24 AHR=67.11	UHR=45.80 VHR=86.45 AHR=66.12	UHR=46.28 VHR=88.74 AHR=67.51	UHR=47.40 VHR=89.24 AHR=68.32
Babble (0 dB)	UHR=47.20 VHR=85.42 AHR=66.31	UHR=46.52 VHR=84.86 AHR=65.69	UHR=47.00 VHR=86.28 AHR=66.64	UHR=48.20 VHR=87.42 AHR=67.81
Train (0 dB)	UHR=45.26 VHR=84.86 AHR=65.06	UHR=44.28 VHR=86.24 AHR=65.26	UHR=46.48 VHR=87.00 AHR=66.74	UHR=47.86 VHR=87.56 AHR=67.71



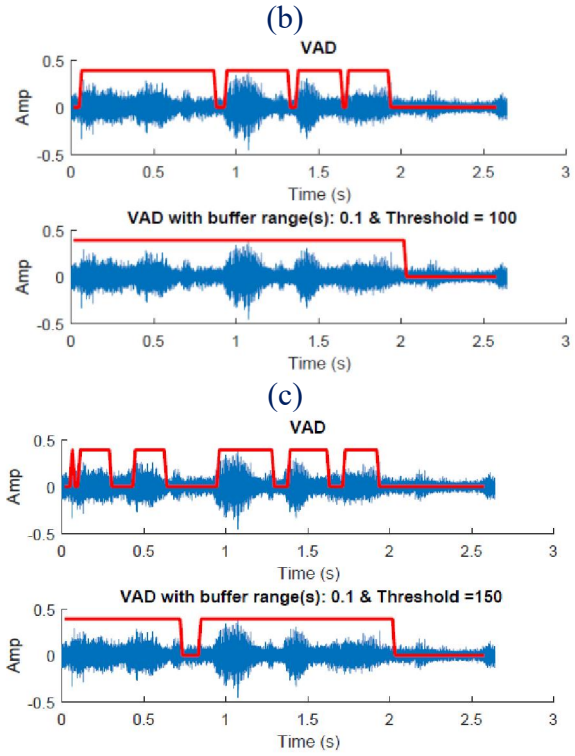


Figure 5: Spectral energy and corresponding VAD decision for threshold values of (b) 50, (c) 100 and (d) 150 for train noise at 0 dB SNR level.

Conclusion and Future Work

In this work, an extensive study with different features for VAD was performed. Results indicated that the spectral energy based VAD with adaptive threshold were useful for low SNR signal. Future work shall include exploring the combination of features for robust VAD for negative SNR level.

Acknowledgements

This work was supported by Dept. of E&CE, Jain Institute of Technology, Davangere and Visvesvaraya Technological University, Belagavi.

References

1) Drugman, T., Stylianou, Y., Kida, Y., Akamine, M. (2016). Voice Activity Detection: Merging Source and Filter-based Information. IEEE Signal Processing Letters. 23(2):252–256.
2) Benyassine, A., ITU-T Recommendation G. (1997). 729 Annex B: A silence compression scheme for use with G. 729 optimized for V. 70 digital simultaneous voice and data applications. IEEE Communication Magazine 35:64–73.
3) Tong, A., Chen, N., Qian, Y. Yu, K. (2014). Evaluating VAD for automatic speech recognition. Proc. IEEE Int Conf. on Signal Processing, IEEE, Hangzhou, China, 2014:2308–2314.

4) Moattar, M.H., Homayounpour, M.M. (2009). A simple but efficient real-time voice activity detection algorithm. European Signal Processing Conference,Eusipco. 10:2549–2553.
5) Kim, S.K., Kang, S.I., Park, Y.J., Lee, S., Lee, S. (2016). Power spectral deviation-based voice activity detection incorporating teager energy for speech enhancement. Symmetry. Basel, 8)7):5–12.
6) Pang, J. (2017). Spectrum energy based voice activity detection. IEEE 7thAnnual Computing Communication Working Conference. CCWC 2017, 3:9–13, 2017.
7) Nasibov, Z., Dr. Tomi Kinnunen (2012). Decision fusion of voice activity detectors Keywords. Master's thesis-School of computing, Univ. of Eastern Finland. 1:1-75.
8) Graf, S., Herbig, T., Buck, M., Schmidt, G. (2015). Features for voice activity detection: a comparative analysis. EURASIP Journal of Advanced Signal Processing, 1:1-15.
9) Hu, Y., Loizou, P. (2008). Evaluation of objective quality measures for speech enhancement. IEEE Transactions on Speech and Audio Processing, 16(1):229-238.
10) Ma, J., Hu, Y., Loizou, P. (2009). Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. Journal of the Acoustical Society of America. 125(5):3387-3405.
11) Nautsch, A., Bamberger, B., Busch C. (2016). Decision robustness of voice activity segmentation in unconstrained mobile speaker recognition environments. Lecture Notes Informatics (LNI), Proc. - Ser. Gesellschaft fur Inform., P-2:1-10.