Recent Developments, Challenges, and Future Scope of Voice Activity Detection Schemes—A Review

Shilpa Sharma, Punam Rattan, and Anurag Sharma

Abstract Voice Activity Detection (VAD) is a technique to classify speech signal into two parts as speech signal and background noises, and widely used in emerging speech recognition technologies such as mobile communication, high-quality multimedia transmission, forensic science, and voice recognition applications. As this technique is integral part of speech communication system, selection of precise VAD is the most challenging part in terms of complexity, feature extractions, threshold selection, and percentage of correctness. The researchers have generally classified VAD into supervised and unsupervised system and introduced various characteristics-based algorithm to reflect the occurrence of speech signal. However, a pervasive study is desired for the selection of appropriate techniques from predefined VAD along with the challenges and solutions to set the future research directions in the emerging area of voice recognition. Therefore, an extensive study is presented in this manuscript especially to set a tradeoff between obstacles and performance of earlier developed VAD. The authors believe that this review will be helpful to researchers working in the challenging speech processing and recognition domain.

Keywords VAD · Speech recognition · GMM · HMM · GSM

S. Sharma (✉)
Department of CSE, CT University, Ludhiana, India

P. Rattan
Department of Computer Application, CT University, Ludhiana, India

A. Sharma
Department of Computer Science and Engineering, GNA University, Phagwara, India
e-mail: er.anurags@gmail.com

M. S. Kaiser et al. (eds.), *Information and Communication Technology for Competitive Strategies (ICTCS 2020)*, Lecture Notes in Networks and Systems 190,
https://doi.org/10.1007/978-981-16-0882-7_39

1 Introduction

The fundamental task of Voice Activity Detection (VAD) is two poles apart the noise and audio segments from the speech signal. This field has immensely used in various fields such as speech improvement, speech coding [1] speech surveillance, speech recognitions [2], and language identification [3]. Therefore, the major expectation from VAD algorithm is high effectiveness along with robustness to noise and lesser computational complexity. In this direction, various studies and investigations have been reported in previous work. Broadly, entire work is categorized in two steps: (a) discrimination model and (b) feature extraction. Initially, more efforts were focused on the integration of energy-based features with zero-crossing rate (ZCR) approach [2, 3] which is highly affected by additive noise. To mitigate the additive noise affects, several other approaches have been adopted by researchers such as Mel-Frequency Cepstral Coefficients (MFCCs) [4], ancestral distance [5], autocorrelation-based features [6], line spectral frequencies [7], linear prediction (LP) residual [8] and periodicity-based features [9]. Besides these methods, discrete Fourier transforms (DFT) based statistical models have also been proposed [10, 11]. However, few studies have explored the variable properties of speech and noise segments in an audio signal [12, 13] while hybrid approaches have been cited in [14, 15]. Thereafter, artificial intelligence came into picture [15, 16] to mitigate the limitation of traditional speech recognition systems which use Gaussian Mixture Models (GMMs) based Hidden Markov models (HMMs) to represent a speech segment. This conventional method gained popularity due to its simple design and its practical design but is an inefficient approach for nonlinear functions [15]. To combat this situation, neural network played a key role, however this scheme better for short time signals and rarely successful for continuous speech signals due to temporal dependencies [16]. The basic diagram for VAD technique is shown in Fig. 1.

This review presents a comprehensive study on the various VAD techniques with overview of speech recognition and comprises of a suitable theoretical background to understand the topic. This manuscript will be helpful to identify research models in previous years along with the research gaps and limitations as well as the future

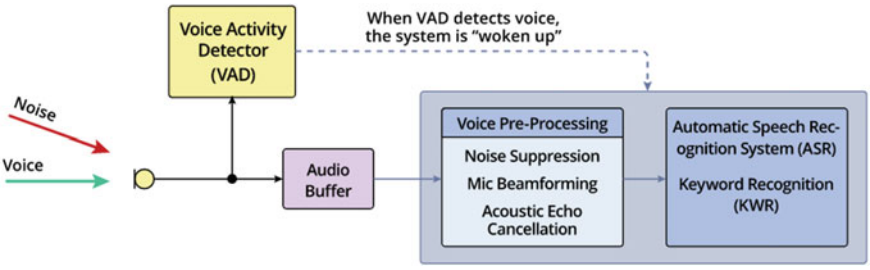

Fig. 1 Block diagram of voice activity detection [39]

outlines. Authors believe that this review will help future researcher to identify new research directions to fill the gap in existing VAD approaches.

2 Background

The major area of VAD is speech recognition which contains various types of feature such as speaker recognition emotion recognition, health recognition, language recognition, accent recognition, age recognition, and gender recognition. The design overview of automatic speech recognition system is shown in Fig. 2. The speech recognition has a wide variety of applications in the area of computers dictation instead of typing, spaceships, handicapped people, Smart home, and many more. Generally, automatic speaker based approaches are divided into two processes as speaker identification and speaker verification. In speaker identification process identifies the registered speaker and generally use in public facilities/media. The procedure for admitting and discarding the identified speaker is known as speaker verification. The applications of verification process are healthcare, remote access to computer, dataset access, and telephone networks [17–21]. The other attractive domain of speaker recognition is emotion cue based speaker identification. As human computer based technique is maturing day by day, hence emotion based speaker recognition has capability of affective computing based speech recognition and gained popularity. The machine based emotion recognition defines the task of recognition of unknown emotion based information within the speech signals. Further, this technique is divided into two parts as emotion identification and emotion recognition. The applications of emotion recognition are stated in [22, 23]. On the other hand, to update the patient health status using their voice is known as automatic health recognition system. Additionally, speech recognition system has a smart application in the area of language recognition where language is identified and facing the critical challenge when language is closely correlated. Another emerging area of VAD in speech recognition is age estimation along with gender identification based on audio samples. Mostly, VAD algorithms are divided into two parts; feature extraction and detection schemes.

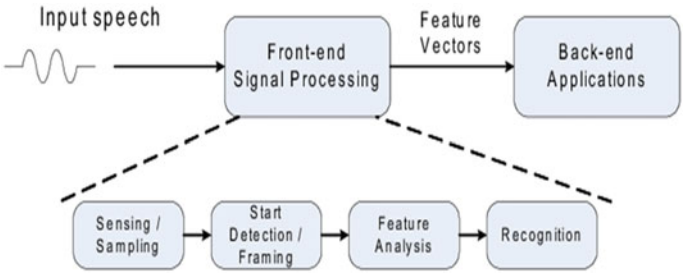

Fig. 2 Design of automatic speech recognition system [18]

3 Development in VAD

Initially, the GSM 729 [23] defined the VAD module for variable bit speech coding but less robust to noisy environment. The robustness to noise is also an important step as it can improve the performance of automatic speech recognition in noisy environment. According to the state of art in voice activity detection, many algorithms have been proposed. Most of these proposed algorithms are differentiated on the basis of features incorporated. Out of all features, short term energy and zero crossing rate has been gained more popularity due to their less complexity but are not effective in noisy environment. Therefore, acoustic features came into picture to combat the problem of Short term energy and zero crossing rate. The main acoustic feature based developments in VAD are auto-correlation based function [4], Mel-frequency [24], power in the band limited region [7], delta line spectral frequencies [25] and features-based higher-order statics. All these developments took place during 2000–2005. All these experiment improved the robustness against diverse environments with less complexity and elevated efficiency. Meanwhile, combination of multiples features based VAD algorithms have also been tested, for instance, CART [26], ANN [27] but augmented the complexity. Additionally, some efforts have been initialized for noise characterization, hence enhanced speech spectra derived from Wiener filtering based noise statistics [28] came in existence. These methods are more sensitive to variation in SNR as they developed with assumption of stationary noise. Some other works have also reported noise estimation and adaptation to improve the robustness in VAD with additional computational complexity. Additionally, ETSI AMR [29] and the AFE [30], have been used as VAD techniques for various proposed algorithms. As the applications of VAD are increasing, some new features are introducing such as wavelet transformed, spectral entropy, and wavelet energy entropy ratio [31, 32]. These features have easiness in realization and more appropriate for random signal processing with a limitation to detect end point of speech signal efficiently under noisy environment. Further, to enhance the decision performance, a power spectral density based Teager Energy (TE) is derived. Even IS-127 emerged as feature of detection in VAD [33].

Recently, the unsupervised methods attracted many researchers to improve the unsupervised VADs, such as GMM-based VAD [34]. This method has used integration of log-likelihood ratio and short time energy based voice activity detection feature. Meanwhile, a self-adaptive VAD based on vector quantization scheme is proposed in [35]. Besides that, the deep multimodal end to end architecture, visual and audio network [36], diffusion networks based network towards transient noise [37, 38] emerged as recent developments in the domain of VAD methods. In 2017, OpenSAT was started by National Institute of Standard and Technology and prolongation is reported by name of OpenSAT 2019. The OpenSAT is planned to keep development in the domain of speech Activity detection (SAD), Keyword search, and ASR. However, the primary focus of OpenSAT'19 is to develop Public Safety Communication (PCS), Video Annotation for Speech Technology (VAST), and Low

Table 1 Comparison of VAD techniques

Technique	Pros	Cons	Solutions
Hidden Markov models (HMMs) based Gaussian Mixture	Simple design and its practical design	Inefficient approach for nonlinear functions	Neural network based VAD
Short Term Energy And Zero Crossing Rate	Less complexity	Ineffective in noisy environment	Auto-correlation based function, Mel-frequency, delta line spectral frequencies, and features-based higher order statics
CART, ANN	Effective in noisy environment	Augmented the complexity	Wiener filtering
Wavelet Transformation, Spectral Entropy, and Wavelet Energy Entropy Ratio	Easiness in realization and more appropriate for random signal processing	Limitation to detect end point of speech signal efficiently under noisy environment	A power spectral density based Teager Energy (TE)
GMM-based VAD	Integration of log-likelihood ratio and short time energy based voice activity detection feature	Unsupervised VADs only	Self-adaptive VAD based on vector quantization scheme
Multimodal VAD	Highly effective in noisy and transient environment	Incorporation of the video signal	Multimodal Compact Bilinear Pooling (MCBP)

Resourced Language (LRL). Furthermore, new scheme has been proposed as multimodal VAD which is highly effective in noisy and transient environment. However, there are numbers of techniques and research efforts have been in initiated, even some of them are matured enough or near to mature still there are many challenges and opportunities in domain of VAD. A comparative analysis of most popular VAD technique along with pros and cons is presented in Table 1. In the next section, the outlines for future research direction along with challenges and opportunities have been highlighted.

4 Challenges and Opportunities

The various VAD architectures and algorithms have been developed to improve the robustness to noisy environment as it is proposed as the most critical issue since there are wide varieties of noises which behave differently and can significantly affect the performance of any algorithm. One of the widely used platform is NOISE-92 which

still does not have a good set of noise type to known real-world noise natures. Therefore, huge opportunities are available for audio classification framework which is key demand for VAD, hence needed to be tackled at present. Further, multimodal VAD is at their earlier stage of development, and need to be focused as it comprised of audio and video features. The surprising fact observed during previous work is that most of the academician and researchers are still using MFCCs as feature extraction technique for voice signal in deep learning environment. Conversely, MFCCs is a classical classifier, hence it is necessary to implement other feature extraction methods such as Linear Predictive Coding (LPC) especially in deep learning environment. This is the most interesting and challenging research area in the domain of VAD. In a study, it is mentioned that about 75% of DNN models are standalone where nearly 25% focus on hybrid techniques. Therefore, there is a huge scope for researcher and professionals of VAD domain to use hybrid models as few initiated efforts in this direction showed promising outcomes. Hence authors encourage analyzing hybrid model technique to improve the robustness to noisy environment in VAD architectures. Additionally, it is observed that very few research efforts are kicked off in the direction of Recurrent Neural Network (RNN). Besides RNN, Long Short Time Memory (LSTM) can be proven as a powerful method to improve the VAD and speech recognition. As computational complexity of any algorithm play an important role, hence to reduce the complexity of already available approaches can be seen as future direction, even if many attempts have already been made. Nowadays, artificial intelligence, deep learning, and machine learning approaches are gaining popularity in many domains. The implementation of these technologies in VAD is premature which needed to tested to improve existing VAD methods. Furthermore, most of the work in VAD is based on statistical model and out of which very less VAD architectures are tested under real-world scenario. Therefore, professional can work in direction to implement and test these statistical models in real-time environment.

5 Conclusion

The manuscript presented a comprehensive review of the various methods and techniques in the domain of VAD along with the development, challenges and opportunities. The information will be helpful to the professional of VAD domain to choose a research area to improve the robustness against noisy environments and VAD. VAD is integral part of speech communication system; therefore, selection of precise VAD is most challenging part in terms of complexity, feature extractions, threshold selection and percentage of correctness. Therefore, various feature extortion techniques are cited along with limitation, advantages, and solutions. The authors believe that this widespread proportional study will be helpful to researcher working in the challenging speech processing and recognition domain.

References

1. I. Mc Cowan, D. Dean, M. McLaren, R. Vogt, S. Sridharan, The delta phase spectrum with application to voice activity detection and speaker recognition. IEEE Trans. Audio Speech Lang. Proc. **19**, 2026–2038 (2011)
2. D. Valj, B. Kotnik, B. Horvat, Z. Kacic, A computationally efficient mel filter bank VAD algorithm for distributed speech recognition systems. Eurasip J. Appl. Sig. Process. **4**, 487–497 (2005)
3. B. Kotnik, Z. Kacic, B. Horvat, A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm, in *Proceedings of 7th Europseech* (2001), pp. 197–200
4. T. Kristjansson, S. Deligne, P. Olsen, Voicing features for robust speech detection, in *Proceedings of Interspeech* (2005), pp. 369–372
5. J. Haigh, J. Mason, A voice activity detector based on Cepstral analysis, in *Proceedings of Eurospeech* (2003), pp. 1103–1106
6. S.O. Sadjadi, J. Hansen, Unsupervised speech activity detection using voicing measures and perceptual spectral flux. IEEE Sig. Pro. Lett. **20**, 197–200 (2013)
7. M. Marzinzik, B. Kollmeier, Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. IEEE Trans. Speech Audio Process. **10**, 109–118 (2002)
8. E. Nemer, R. Goubran, S. Mahmoud, Robust voice activity detection using higher-order statistics in the LPC residual domain. IEEE Trans. Speech Audio Process. **9**, 217–231 (2001)
9. K. Ishizuka, T. Nakatani, Study of noise robust voice activity detection based on periodic component to aperiodic component ratio, in *Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition* (2006), p. 6570
10. J. Ramirez, J. Segura, M. Benitez, L. Garcia, A. Rubio, Statistical voice activity detection using a multiple observation likelihood ratio test. IEEE Sig. Proc. Lett. **12**, 689–692 (2005)
11. P. Ghosh, A. Tsiartas, S. Narayanan, Robust voice activity detection using long-term signal variability. IEEE Trans. Audio Speech Lang. Process. **19**, 600–613 (2011)
12. Y. Kida, T. Kawahara, Voice activity detection based on optimally weighted combination of multiple features, in *Proceedings of Interspeech* (2005), pp. 2621–2624
13. S. Soleimani, S. Ahadi, Voice activity detection based on combination of multiple features using linear/kernel discriminant analyses, in *Proceedings of Information and Communication Technologies: From Theory to Applications* (2008), pp. 1–5
14. H. Singh, A.K. Bathla, A survey on speech recognition. Int. J. Adv. Res. Comput. Eng. Technol. **2**(6), 2186–2189 (2013)
15. M.A. Anusuya, S.K. Katti, Speech recognition by machine: a review. Int. J. Comput. Sci. Inf. Secur. **6**(3), 181–205 (2009)
16. J. Padmanabhan, M.J.J. Premkumar, Machine learning in automatic speech recognition: A survey. IETE Tech. Rev. **32**(4), 240–251 (2015)
17. C.-C. Shen, W. Plishker, S.S. Bhattacharyya, Design and optimization of a distributed, embedded speech recognition system, in *Proceedings of the International Workshop on Parallel and Distributed Real-Time Systems*, Miami, Florida, April 2008
18. G. Zhou, J.H.L. Hansen, J.F. Kaiser, Nonlinear feature based classification of speech under stress. IEEE Trans. Speech Audio Process. **9**(3), 201–216 (2001)
19. C. Fredouille, G. Pouchoulin, J.-F. Bonastre, M. Azzarello, A. Giovanni, A. Ghio, Application of automatic speaker recognition techniques to pathological voice assessment (dysphonia), in *Proceedings of 9th European Conference on Speech Communication and Technology (Eurospeech)* (2005), pp. 149–152
20. V.A. Petrushin, Emotion recognition in speech signal: experimental study, development, and application, in *Proceedings of Sixth International Conference on Spoken Language Processing (ICSLP)* (2000), p. 5
21. N. Fragopanagos, J.G. Taylor, Emotion recognition in human–computer interaction. Neural Netw. **18**(4), 389–405 (2005)

22. E. Douglas-Cowie, N. Campbell, R. Cowie, P. Roach, Emotional speech: towards a new generation of databases. Speech Commun. **40**(1–2), 33–60 (2003)
23. B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, R. Sarikaya, Robust speech recognition in noisy environments: The 2001 IBM SPINE evaluation system. Proc. ICASSP **1**, 53–56 (2002)
24. ETSI standard document, ETSI ES 202 050 V 1.1.3. (2003)
25. K. Li, N.S. Swamy, M.O. Ahmad, An improved voice activity detection using higher order statistics. IEEE Trans. Speech Audio Process. **13**, 965–974 (2005)
26. G.D. Wuand, C.T. Lin, Word boundary detection with MEL scale frequency bank in noisy environment. IEEE Trans. Speech Audio Process. (2000)
27. A. Lee, K. Nakamura, R. Nisimura, H. Saruwatari, K. Shikano, Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs, in *Interspeech* (2004), pp. 173–176
28. B. Lee, M. Hasegawa-Johnson, Minimum mean squared error a posteriori estimation of high variance vehicular noise, in *Proceedings of Biennial on DSP for In-Vehicle and Mobile Systems*, Istanbul, Turkey, June 2007
29. ETSI ES 202 050 Recommendation, Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms (2002)
30. C.F. Juang, C.N. Cheng, T.M. Chen, Speech detection in noisy environments by wavelet energy-based recurrent neural fuzzy network. Exp. Syst. Appl. **36**(1), 321–332 (2009)
31. K.C. Wang, Y.H. Tasi, Voice activity detection algorithm with low signal-to-noise ratios based on spectrum entropy, in *Second International Symposium on Universal Communication (ISUC'08)* (2008), pp. 423–428
32. S.K. Kim, S.I. Kang, Y.J. Park, S. Lee, S. Lee, Power spectral deviation-based voice activity detection incorporating teager energy for speech enhancement. Symmetry **8**(7), 58 (2016)
33. F.G. Germain, D.L. Sun, G.J. Mysore, Speaker and noise independent voice activity detection, in *Interspeech* (2013), pp. 732–736
34. T. Kinnunen, P. Rajan, A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data, in *ICASSP* (2013), pp. 7229–7233
35. I. Ariav, I. Cohen, An end-to-end multimodal voice activity detection using wavenet encoder and residual networks. IEEE J. Sel. Topics Sig. Process. **13**(2), 265–274 (2019)
36. A. Ivry, B. Berdugo, I. Cohen, Voice activity detection for transient noisy environment based on diffusion nets. IEEE J. Sel. Topics Sig. Process. **13**(2), 254–264 (2019)
37. H. Dubey, A. Sangwan, J.H. Hansen, Leveraging frequency dependent kernel and dip-based clustering for robust speech activity detection in naturalistic audio streams. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(11), 2056–2071 (2018)
38. G.-B. Wang, W.-Q. Zhang, An RNN and CRNN based approach to robust voice activity detection (2019). https://doi.org/10.1109/apsipaasc47483.2019.9023320
39. Available online http://www.alango.com/voice-activity-detection.php