Audio enhancement using modified spectral subtraction approach through adaptive noise estimation using Support Vector Machine based Voice Activity Detection

Shambhu Shankar Bharti
Assistant Professor, Lok Nayak Jai Prakash Institute of Technology, Chapra
Rajeev Ranjan
Assistant Professor, Bakhtiyarpur College of Engineering, Bakhtiyarpur
Bhawana Singh
Assistant Professor, Bakhtiyarpur College of Engineering, Bakhtiyarpur
Ajeet Kumar
Assistant Professor, Bakhtiyarpur College of Engineering, Bakhtiyarpur

Abstract:
Improving the perceptual quality of audio signal in terms of degree of listener fatigue, intelligibility is the primary need of any audio based smart devices. Audio enhancement is needed for the applications like voice biometrics, audio-to-text conversion etc. Number of stationary and non-stationary background noises adds with the audio signal while recording or transmitting them. Addition of background noise signal deteriorates the quality of audio which decreases the efficiency of audio based applications. Various variants of spectral subtraction methods have been used for enhancing the quality of audio signal by reducing the background noises. This paper enhances the quality of audio by estimating the noise followed by applying modified spectral subtraction method. Here, noise is updated by identifying the absence of human speech signal in a frame.

Keywords: Plant diseases, AI, computer vision, disease detection, neural networks, support vector machines, k-nearest neighbors, agriculture.
DOI Number: 10.14704/nq.2018.16.10.1693 **NeuroQuantology 2018; 16(10):14-21**

1. Introduction

Audio is the most common and easiest way of communication among human beings. In today`s busy life, human communicates with others via mobile phone/ Walkie Talkie and other wired/wireless devices in public places. Generally, background noise gets added with the audio signal and communicates via the communicated medium which degrades the transmitted audio quality. Many times, background noise is so prominent that the actual sound is not understood by the listener. Thus, preprocessing and enhancement of the audio signal are required before its gets transmitted in real life. Audio enhancement is also essential for Hearing-Aid devices. Hearing-Aid devices are worn by a person with hearing loss to speak, listen, and engage more fully in everyday activities by amplifying some sounds. There are many other applications of audio

enhancement including speech recognition system, speaker recognition/identification system, telecommunications, voice communications and many more.

Enhancing certain perceptual aspects like clarity, quality and intelligibility of the audio signal is the main focus of audio enhancement techniques. In other words, generally noise is suppressed from the audio signal to enhance the quality, clarity and intelligibility of audio signal. There are many types of noise signals like background noise, electrical noise and many more are added during the recording/transmission of the audio. Therefore, estimation of noise signal plays an important role for any audio enhancement techniques. Audio enhancement can be made either in the frequency domain, temporal domain, or in combination of these. Spectral subtraction based approach, Statistical based approach, Subspace based approach, and auditory masking based approaches are some of the approaches that have been used for enhancing the audio signal by researchers.

In this paper, we are proposing Audio enhancement using modified spectral subtraction approach through adaptive noise estimation using Support Vector Machine based Voice Activity Detection. This work is mainly focused to enhance the quality of audio by suppressing the additive noise present in the audio signal. For, this we have assumed that the portion of the audio which does not contain human speech is taken as noise signal. The next section presents the background of the earlier state-of-the-art approaches in the field of speech enhancement.

2. Background

Degraded audio can be made more comprehensible and of higher quality through audio enhancement. Estimating noise in an audio enhancement approach is a crucial and difficult task. Voice Activity Detector (VADr) can be used to estimate the noise signal. The parts that contain human speech are identified by VADr; the other portions that are not recognized by VADr may be interpreted as noise. This section is a review of the research on voice activity detection techniques and how they are used to improve audio.

2.1 State-of-the-Art techniques of Voice Activity Detection

Numerous investigations have employed several acoustic characteristics, including spectral correlation and energy entropy, within the domain of vocal activity detection (VAD). The vocal tract properties and the excitation sequence are the two main pieces of information in a human audio signal. Voice excitation sequence features have been the only ones used by several researches for VAD [1]. Energy entropy, low-frequency ultrasound, single frequency filtering, spectrum divergence, spectral correlation, and other properties have been used for the same purpose [2], [3], [4], and so on. Long-term or suprasegmental features, such as long-term spectral divergence measure (LTSDM) and long-term signal variability (LTSV), have been used in a number of research on VAD [5]. These studies have shown that these features function well even at low SNRs. Over longer duration samples, LTSDM examines the spectral divergence between human speech and noise [6]. For the aforementioned goal, features with spectral energy in various frequency bands and scales, such as the bark-scale or MEL-scale, have been widely exploited. Mel Frequency Cepstral Coefficient (MFCC) is the term used to refer to spectral energy in MEL-Scale [7]. Many researchers have collected and employed features from voiced sounds (exhibiting quasi-periodic behavior and dynamic spectrum characteristic for short utterance) for VAD [8], [9], [10].

Numerous statistical techniques are also widely used in the literature, including the log-likelihood ratio test [11], the optimum likelihood ratio test [12], the low-variance spectrum estimate [13], the Laplace distribution [14], the Gaussian distribution [15], the Gamma distribution [16], and the combination of them [17]. These techniques have excellent computational efficiency, but as SNR values drop, their systems' accuracy sharply declines as well. These voice activity detectors only use a

very small number of variables that capture particular aspects of human speech. Performance is inconsistent as a result. Combining several elements could make VAD operate better. Multiple feature fusion is used for VAD in several machine learning techniques, such as artificial neural networks [18], Support Vector Machine (SVM) [19], [20], Deep Belief Network (DBN) [21], etc.

A popular and effective method for VAD is the machine learning-based technique. There are two types of approaches for this: supervised and unsupervised. Dimensionality reduction on the retrieved feature is done initially in an unsupervised manner. With enough labeled dataset, supervised trained classifiers for VAD work effectively in nearly all noise conditions.

2.2 State-of-the-art Techniques for Speech Enhancement using Spectral Subtraction Method

The spectral-subtraction technique is among the earliest algorithms for speech enhancement that have been proposed in history. One expects additive noise when using this strategy. The estimated noise spectrum is subtracted from the noisy signal spectrum to provide an enhanced signal spectrum. Enhanced speech is then created by performing the inverse Fourier transform on the enhanced signal spectrum. The use of spectral subtraction for the suppression of acoustic noise in speech was suggested by S.F. Boll [22]. He evaluated the noise during non-speech activities in this paper, on the assumption that the noise would either be stationary or vary extremely slowly. The approach requires a voice activity detector for the slowly varying non-stationary noise in order to inform the program when speech has stopped and to estimate a new noise bias. It is believed that by eliminating the impact of noise from the magnitude spectrum alone, considerable noise reduction is achievable. A non-linear spectral subtraction was proposed by P. Lockwood et al. [23]. They noticed that certain noise can have a greater impact on the low-frequency than the high-frequency portions of the spectrum. The proposal put out was to subtract bigger values at frequencies with low signal-to-noise ratios and smaller values at frequencies with high SNR ratios. For listeners with hearing impairments, S. K. Waddi et al. [24] suggested a speech enhancement method based on cascaded-median based noise estimation and spectrum removal. In order to minimize memory usage and computational cost, it estimates the noise spectrum using a cascaded-median approach.

3. Proposed Approach for Audio Enhancement using adaptive noise estimation through Voice Activity Detection

3.1 Noise Identification System

The block diagram for the suggested noise identification system is displayed in Figure 1. To do this, VAD G.729 is used to extract an appropriate set of noise frames. It is discovered through experimentation that the boundaries (the beginning or end of the audio frame where human speech is heard) are where the misclassification rate for VAD increases. In order to reduce it, each i^{th} frame is considered a noise frame if the $(i-1)^{th}$, $(i-2)^{th}$, and $(i+1)^{th}$ frames are all identified by VAD G.729 as noise frames. The Random Forest classifier is then trained using the LPC and MFCC features that were previously retrieved from these noise frames, allowing it to recognize the different types of noise.

3.2 Noise Intensity Estimation

Voice Activity Detection is used to identify the presence or absence of human voice in each frame. It is presumed that a frame only contains noise if human speech is not recognized inside it. The noise database is updated using additional noise frames. The ultimate noise intensity is determined by taking the average of previously identified 200 noise frames. Then, to improve the quality of this frame, modified spectral subtraction is applied, as proposed by Bharti et al [25]. When human speech is identified in a frame, the modified spectral subtraction approach is applied to enhance the current frame by using the previous updated noise intensity as its current noise intensity. Thus, by improving every frame, hidden audio is improved.

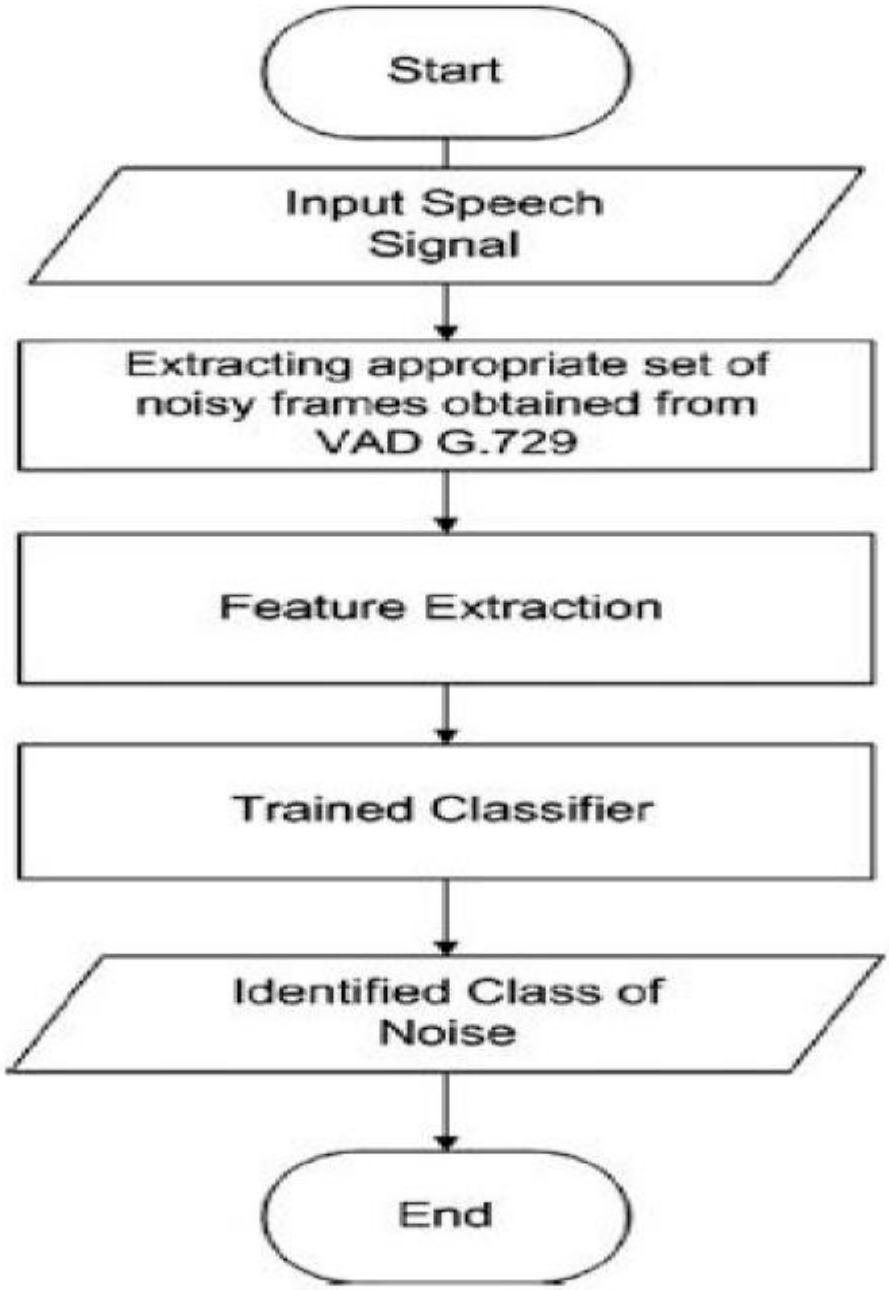

Fig.1: Noise Identification System

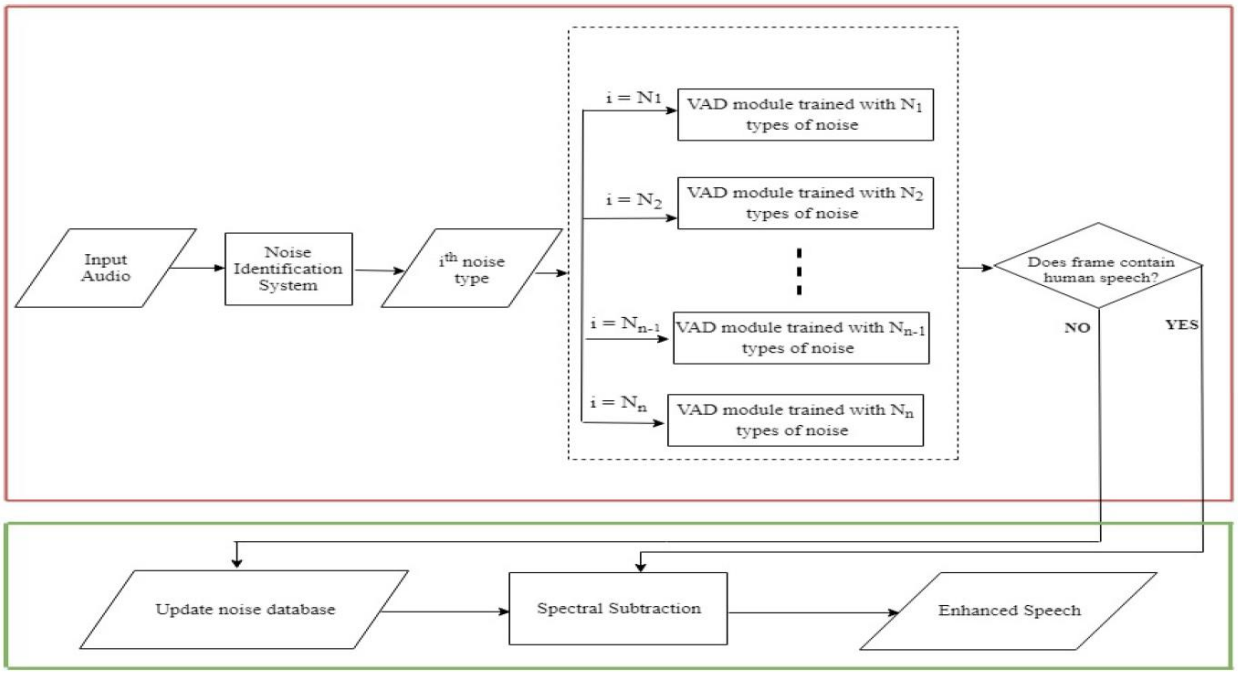

Fig.2: Noise Estimation Block diagram System

4. Experimental Results and their Analysis

Using the NOIZEUS [26] database, studies have been carried out to evaluate the efficacy of the suggested strategy. Thirty IEEE clear sentences are in it, and three male and three female speakers deliver them. Eight sounds from the actual world have also been combined with these lines, at four different SNRs (0dB, 5dB, 10dB, and 15dB): airport, babble, car, exhibition hall, restaurant, railway station, street, and train. Additional noises were extracted from the AURORA database. Setting parameters The most often utilized objective/subjective metrics, SNR and MOS, have been employed to gauge the improved audio's quality. Fifteen people are asked to score the clarity, quality, and intelligibility of the upgraded audio in comparison to the noisy original audio in order to determine the MOS. The following numerical values are assigned: 1, 2, 3, and 4 for bad, poor, fair, and good, respectively. The average rating for each audio is used to calculate the MOS.

Table 1 shows the SNR value of noisy audio and enhanced audio when noise is updated using the proposed techniques.

Table1: SNR Value of noisy audio and enhanced audio

Type of noise	SNR (dB)			MOS
	noisy audio	enhanced audio	Improvement	
Babble noise	0	6.87	6.87	3
Babble noise	5	9.24	4.24	3.35
Babble noise	10	13.76	3.76	3.50
Babble noise	15	16.52	1.52	3.7
Car noise	0	6.49	6.49	3.2
Car noise	5	8.47	3.47	3.3
Car noise	10	12.48	2.48	3.6
Car noise	15	16.47	1.47	3.5
Exhibition noise	0	5.78	5.78	2.80
Exhibition noise	5	8.69	3.69	3
Exhibition noise	10	11.42	1.42	3.25
Exhibition noise	15	17.76	2.76	3.30
Street noise	0	6.93	6.93	3
Street noise	5	8.43	3.43	3.5
Street noise	10	12.65	2.65	3.5
Street noise	15	18.16	3.16	3.5
Restaurant noise	0	6.98	6.98	3.25
Restaurant noise	5	8.15	3.15	3.25
Restaurant noise	10	11.79	1.79	3.5
Restaurant noise	15	17.18	2.18	3
Station noise	0	4.93	4.93	2.40
Station noise	5	7.78	2.78	2.65
Station noise	10	12.35	2.35	3
Station noise	15	17.74	2.74	3.5

By comparing the results presented in Table 1 and results presented in the approach proposed by Bharti et al. [25] following observations are made:

- Enhancement of the objective assessment parameter When using SVM-based VAD instead of energy to update the noise intensity, SNR is higher.
- Evaluation parameter that is subjective When SVM based VAD is used to

update the noise intensity instead of energy, MOS is higher.

Based on the aforementioned findings, it can be said that SVM-based VAD performs better at updating noise intensity than energy feature.

5. Conclusion

In conclusion, it is discovered through experimentation that Audio enhancement using modified spectral subtraction approach through adaptive noise estimation using Support Vector Machine based Voice Activity Detection performs better than adaptive noise estimation using energy features. Additionally, it is also found that improvement in the quality of enhanced audio is well in both stationary and non-stationary noise environments. This method also reduces introduced musical noise during the improvement procedure.

References

1. N. Dhananjaya and B. Yegnanarayana, Voiced/nonvoiced detection based on robustness of voiced epochs, IEEE Signal Process. Lett., 2010, VOL. 17, Issue 3, pp. 273–276.
2. E. Nemer, R. Goubran, and S. Mahmoud, Robust voice activity detection using higher-order statistics in the LPC residual domain, IEEE Transactions on Speech and Audio Processing, 2001, VOL. 9, NO. 3, pp. 217–231.
3. V. McLoughlin, The use of low-frequency ultrasound for voice activity detection, Proc. Interspeech, 2014, pp. 1553–1557.
4. G. Aneeja and B. Yegnanarayana, Single frequency filtering approach for discriminating speech and onspeech, IEEE/ACM Trans. Audio,Speech, Lang. Process., 2015, VOL. 23, NO.4, pp. 705–717.
5. P. Ghosh, A. Tsiartas, and S. Narayanan, Robust voice activity detection using longterm signal variability, IEEE Trans. Acoust., Speech, Language Process., 2011, VOL. 19, NO.3, pp. 600–613.
6. J. Ramrez, J. C. Segura, C. Bentez, A. D. L. Torre, and A. Rubio, Efficient voice activity detection algorithms using long-term speech information, Speech Communication, 2004, VOL. 42, pp. 3–4.
7. T. Kinnunen, E. Chernenko, M. Tuononen, P. Franti and H. Li, Voice activity detection using MFCC features and support vector machine, Proc. Int. Conf. on Speech and Computer (SPECOM07), 2007, VOL. 2, pp. 556–561.
8. I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan, The delta-phase spectrum with application to voice activity detection and speaker recognition, IEEE Trans. Acoust., Speech, Language Process., 2011, VOL. 19, Issue 7, pp.2026–2038.
9. J. Haigh and J. Mason, Robust voice activity detection using cepstral features, Proc. TENCON 1993, pp. 321–324.
10. I-C Yoo, H. Lim and D. Yook, Formant-Based Robust Voice Activity Detection, IEEE/ACM Trans. Audio, Speech, Lang. Process., 2015, VOL. 23, NO. 12, pp. 2238– 2245.
11. J. Sohn, N. S. Kim, andW. Sung, A statistical model-based voice activity detection, IEEE Signal Process. Letters, 1999, VOL. 6, NO. 1, pp. 1–3.
12. J. Ramirez, J. C. Segura, C. Benitez, L. Garcia and A. Rubio, Statistical voice activity detection using a multiple observation likelihood ratio test, IEEE Signal Processing Letters, 2005, VOL. 12 Issue 10, pp. 689–692.
13. 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, 2006, VOL. 14, Issue 2, pp. 412–424.
14. S. Gazor and W. Zhang, A soft voice activity detector based on a Laplacian-Gaussian model, IEEE Trans. Acoust., Speech, Language Process., 2003, VOL. 11,NO. 5, pp. 498–505.
15. J. Sohn, N. S. Kim, andW. Sung, A statistical model-based voice activity

detection, IEEE Signal Process. Letters, 1999, VOL. 6, NO. 1, pp. 1–3.

16. J. H. Chang, N. S. Kim and S. K. Mitra, Voice activity detection based on multiple statistical models, IEEE Transactions on Signal Processing, 2006, VOL. 54, NO. 6, pp. 1965–1976.
17. Y. Ephraim and D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator, IEEE Trans. Audio, Speech, Signal Process, 1984, VOL. 32, NO. 6, pp. 1109–1121.
18. T. Pham, C. Tang, and M. Stadtschnitzer, Using artificial neural network for robust voice activity detection under adverse conditions, Int. Conf. on Computing and Communication Technologies, RIVF, 2009, pp. 1–8.
19. T. Kinnunen, E. Chernenko, M. Tuononen, P. Franti and H. Li, Voice activity detection using MFCC features and support vector machine, Proc. Int. Conf. on Speech and Computer (SPECOM07), 2007, VOL. 2, pp. 556–561.
20. J. Ramirez, P. Ye lamos, J.M. Gorriz and J.C. Segura, SVM-based speech endpoint detection using contextual speech features, Electronics Letters, 2006, VOL. 42, Issue 7, pp. 426–428.
21. X.-L. Zhang and J. Wu, Deep belief networks based voice activity detection, IEEE Trans. Acoust., Speech, Language Process., 2013, VOL. 21, NO. 4, pp. 697–710.
22. S.F. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Tans. Acoust., Speech, Signal Processing, April 1979, VOL. ASSP-27, pp.113–120.
23. P.Lockwood and J. Boudy. Experiments with a Non-linear Spectral Subtractor(NSS), Hidden Markov Models and the projections, for robust recognition in cars, Speech Communication., 1992, pp. 215–228.
24. S. K.Waddi, P. C. Pandey, and N. Tiwari, Speech Enhancement Using Spectral Subtraction and Cascaded-Median Based Noise Estimation for Hearing Impaired Listeners, IEEE Conference, 2013.
25. S.S. Bharti, M. Gupta, and S. Agarwal, A new spectral subtraction method for speech enhancement using adaptive noise estimation. In *2016 3rd international conference on recent advances in IEEE information technology (RAIT)* pp. 128-132, 2016.
26. Y. Hu and P. Loizou, Subjective evaluation and comparison of speech enhancement algorithms, Speech Communication, 2007, VOL. 49, pp. 588–601.