Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19174-z

An ensemble of optimal smoothing and minima controlled through iterative averaging for speech enhancement under uncontrolled environment

Raghudathesh G P[1] · Chandrakala C B[2] · Dinesh Rao B[3] · Thimmaraja Yadava G[4]

Received: 14 June 2023 / Revised: 4 March 2024 / Accepted: 2 April 2024
© The Author(s) 2024

Abstract
Although better progress has been made in the area of speech enhancement, a significant performance degradation still exists under highly non-stationary noisy conditions. These conditions have a detrimental impact on the performance of the speech processing applications such as automatic speech recognition, speech encoding, speaker verification, speaker identification, and speaker recognition. Therefore, in this work, a robust noise estimation technique is proposed for speech enhancement under highly non-stationary noisy scenarios. The proposed work introduces an optimal smoothing and minima controlled (OSMC) through an iterative averaging method for noise estimation. Firstly, the computation of smooth power spectrum of degraded speech data and tracking the minima by continuously taking the past spectral average values are considered. Then, to find the activity of speech in each frequency bin, the ratio of degraded speech spectrum to its local minimum is considered, and a Bayes minimum-cost rule is applied for the decision-making. Finally, the spectrum of noise is estimated using the time-frequency dependent smoothing factors which mainly depend on the estimation of the probability of speech presence. The experiments are conducted on NOIZEUS and Kannada speech databases. The evaluated results demonstrated that the proposed OSMC technique exhibits better speech quality and intelligibility performance compared to existing algorithms under highly non-stationary noisy conditions.

Keywords Speech enhancement · Noise estimation · Optimal smoothing and minima controlled (OSMC) · NOIZEUS and Kannada speech databases · Speech quality · Speech intelligibility

Chandrakala C B, Dinesh Rao B and Thimmaraja Yadava G are contributed equally to this work.

✉ Chandrakala C B
chandrakala.cb@manipal.edu

Extended author information available on the last page of the article

Published online: 17 April 2024

1 Introduction

The process of speech enhancement plays an important role in performance of the speech processing applications under real time conditions. The speech data corrupted by various types of noises will have an adverse effect on the performance of the system. Therefore, it is indeed to reduce the various types of noises in the degraded speech data using noise reduction techniques [1]. A technique based on optimal smoothing and minimum-statistics (OSMS) was proposed to calculate the power spectral density (PSD) of non-stationary noise in degraded speech data [2, 3]. The proposed algorithm had the ability to calculate the noise PSD in degraded speech data without using voice activity detection (VAD). Instead of using VAD, the author used tracks of spectral minima in each and every frequency band without having a difference between the activity of speech or speech pause. The optimal smoothing parameter was derived by suppressing the mean square error (MSE) in each step for recursive smoothing of PSD of degraded speech data. The method was amalgamated with a noise reduction algorithm for speech enhancement. The results revealed that the proposed technique outperformed the existing estimators.

The estimation of noise was demonstrated in [4] by proposing a method called minima controlled through recursive averaging (MCRA). By taking the average of past power spectral values of degraded speech data, the noise was approximated. The ratio of local energy of noisy speech to its minimum value within the stipulated time window has given the presence of speech in subbands. The author concluded that the noise approximation is very computationally robust and efficient in terms of signal-to-noise ratio (SNR). The improvement in MCRA for noise approximation under different conditions was demonstrated by Cohen [5]. The noise approximation was calculated by taking the average past power spectral values using a smoothing parameter which was considered by the presence of signal probability. The author claimed that the IMCRA algorithm worked better under non-stationary noisy environments and different SNR conditions.

A noise reduction algorithm was developed in [6] for the speech signals that are degraded by short-time stationary noises and electrical disturbances. The spectral amplitude approximation was used for the enhancement of corrupted speech data. No speech pause detectors were required for the developed noise reduction algorithm, the author claimed. The experimental results showed that there was an improvement in noise reduction in the degraded speech data compared to existing algorithms. The noise estimation and characteristics of noise were explained in [7]. To estimate the noise, the past noisy segments were considered, and no external speech pause detectors were used for the detection of pauses in speech. The implemented algorithm can also be integrated with non-linear spectral subtraction (SS) algorithms, the authors claimed. The results revealed that the proposed algorithm outperformed the existing speech enhancement algorithms. In [8], the time and frequency domain representations of corrupted speech data were used for estimating the noise. The VAD was deployed to identify the active regions of the speech signal. The concept of attenuation procedure was used on speech and non-speech activity regions to obtain the enhanced speech data. The proposed technique had complexity in computation and it is feasible for hearing aid applications. The obtained experimental results were compared with two speech enhancement techniques and the proposed method had given better performance, the authors claimed.

In conclusion, there have been significant advancements in the area of speech enhancement under various degraded conditions. However, research specifically focused on the noise estimation under highly non-stationary noisy conditions is limited [3–10, 19–21, 24]. In the process of noise estimation, understanding the complexities and suppression of various types

of noises under highly non-stationary noisy scenarios would be a unique contribution. Existing works have covered noise estimation techniques under different SNR conditions, but there is a gap concerning the elimination of noise under sudden variations in noise level scenarios. The main objective of the proposed work is to fill this gap by proposing a robust noise estimation technique for speech enhancement under highly non-stationary noisy scenarios. The major contributions to the proposed work are as follows:

- Develop a robust technique for noise estimation in highly non-stationary noisy conditions.
- Estimation of noise in each segment using time-frequency dependent smoothing factors which are calculated using Bayesian probability of speech presence.
- Computation of the noise estimate without using VAD.
- Performance evaluation of the proposed method with existing techniques in terms of speech quality and intelligibility after speech enhancement.

The rest of the article is organized as follows: Section 2 describes the related work. Section 3 gives the implementation of proposed OSMC technique for speech enhancement. The experimental setup and results analysis of proposed and existing noise estimation techniques are described in Section 4. The Section 5 demonstrates the conclusions.

2 Related work

A mono channel speech enhancement algorithm using the Wiener filter was proposed in [21]. The proposed algorithm estimates the noise using a first-order recursive equation and uses a parameter for smoothing to update the noise in every frame. The experiments were conducted on the speech sentences which were degraded by various types of noises. The proposed algorithm outperformed the conventional SS technique in estimating the noise in corrupted speech data, as demonstrated by objective speech quality measures. In [22], a technique was proposed for enhancing a short time spectral amplitude (STSA) in frequency domain. For modeling the magnitudes of DFT of clean speech signal, Weibull distribution was used under Gaussian noise scenario. A decision-directed approach with the smoothing factor was considered to approximate *a priori* SNR. The priori probability of absence of speech was set to 0.2. The experiments were investigated on the TIMIT speech database and the speech sentences were degraded by various additive non-stationary noises. The obtained results demonstrated the efficacy of the proposed technique over existing algorithms in terms of quality and intelligibility of speech.

An algorithm was proposed for noise power spectral density (PSD) estimation using an high-pass filter derivative under noisy conditions [23]. The authors have used a spectral-flatness adaptive thresholding method for the detection of activity of speech in the corrupted speech frames. The NOIZEUS database was used for the subjective and objective evaluation of the proposed noise PSD estimation algorithm and its comparison with competing methods. The achieved results revealed that the proposed technique has performed well for various degraded conditions compared to the competing algorithms. The work in [26] presented a real-time mono channel speech enhancement technique for suppressing stationary and transient noise. The methods employed include quantile noise estimation for stationary noise suppression and the transient noise suppression is based on normalized variance and gravity center of the signal. The results revealed that the proposed algorithm outperformed the existing techniques under high SNR conditions.

The authors have proposed an optimized SS method for mono channel speech enhancement [24]. The proposed method has performed a noise reduction through SS based on

minimal statistics. The obtained results revealed that the proposed method outperformed the competing techniques in terms of quality of speech. In [20], the modified version of minimum-statistics noise estimation (MSNE) was implemented for speech enhancement. The noise was estimated by considering the minima statistics of degraded speech data. The experimental results showed that the proposed estimator outperforms the competing techniques under medium degraded conditions and it has given less performance under highly non-stationary noisy conditions.

An algorithm was proposed for the degraded speech enhancement by combining SS-VAD and MMSE-spectrum power estimator based on zero crossing (SPZC) [9]. The babble and musical noise suppression in degraded speech data was achieved by considering the efficacy of SS-VAD and MMSE-SPZC. For the experimentation, the TIMIT and Kannada speech databases were used. To evaluate the performance of the proposed technique with existing methods, the PESQ and composite measures were considered. The experimental results show that the amalgamation of SS-VAD and MMSE-SPZC estimator has given better improvements in quality and intelligibility of enhanced speech compared to individual techniques. In [25], to optimize the subspace partitioning, the authors have proposed an approach for speech enhancement using modified version of accelerated particle swarm optimization. The degraded speech data was divided into noise only, speech plus noise, and speech only components. The VAD [29–31] was used for the detection of voice activity. The results shown that the proposed method has performed well under different corrupted conditions compared to existing methods. The authors also claimed that there was an improvement with less speech distortion.

The enhancements in the Kannada automatic speech recognition (ASR) system were described in [10]. Initially, the Kannada ASR system was developed in [11] for noisy Kannada speech data. Due to various types of degradation, the performance of the Kannada ASR system led to less speech recognition accuracies. Therefore, the authors in [10] have developed a robust noise reduction technique by combining SS-VAD and MMSE-SPZC estimator [9]. The developed noise reduction technique was integrated with an interactive voice response system as a front end. Further, to improve the accuracy of speech recognition, the deep neural network and subspace Gaussian mixture modeling techniques were used. The experiments were conducted by taking a few speech sentences from TIMIT and Kannada speech databases. The results demonstrated that there was a consistent improvement in the accuracy of speech recognition using the enhanced speech data compared to the earlier Kannada ASR system [11]. A system was developed in [19] for robust speech encoding under degraded conditions. The authors have combined the SS-VAD with LPC for encoding the noisy speech data. The experimental results revealed that the proposed combined SS-VAD and linear predictive coding (LPC) method has given a better performance in terms of SNR and compression ratios. The authors also mentioned that the proposed method works better for higher SNR and is inefficient under medium and negative SNR conditions.

A spatial procedure to SS for speech enhancement was proposed in [27]. The drawbacks of SS were addressed with experimental demonstration of various types of noises. The experimental results revealed that the proposed method has given better results in terms of speech quality and intelligibility compared to existing conventional SS algorithms. In [28], the authors have used the speech enhancement algorithm [27] to improve the accuracy of speech recognition under real-time conditions. The noise elimination algorithm was kept before the feature extraction part in the spoken query system. The experiments revealed that there was an improvement in the suppression of word error rate in the offline ASR models and the accuracy of the speech recognition was improved during the online testing.

In conclusion, the work related to noise estimation has demonstrated significant progress and advancements under different noisy conditions. Nevertheless, one crucial aspect that needs consideration is the estimation of noise under highly non-stationary noisy scenarios. The estimation of noise under an instantaneous increase in noise levels can effectively support improving the efficacy of other speech processing applications such as ASR, speech encoding, speaker verification, speaker identification, speaker recognition and etc,. under noisy conditions.

3 Implementation of proposed OSMC technique for speech enhancement

Consider a sampled degraded speech signal *c(i)* is corrupted by noise model *b(i)* when it is added to clean speech signal *a(i)*, where *i* is the sampling time index. An assumption is made that, *a(i)* and *b(i)* are independent statistically and have zero mean.

$$c(i) = a(i) + b(i) \tag{1}$$

The degraded sampled speech signal *c(i)* is transformed into frequency domain by considering windowing *h(i)* to a frame of *N* consecutive samples of *c(i)*. The fast Fourier transform (FFT) analysis of sliding window yields a set of frequency domain samples (signals) which can be mathematically written as

$$c(\lambda, \mathrm{k}) = \sum_{\mu=0}^{L-1} c(\lambda \mathrm{S} + \mu) h(\mu) \mathrm{e}^{-\mathrm{j}2\pi \mathrm{k}\mu/\mathrm{N}} \tag{2}$$

where λ is subsampled index of time and k is the index of frequency bin and $k \in 0, 1, \ldots, L-1$ which is fairly related to the normalized centered frequency, $\Omega_k = 2\pi k/L$. The probability density function (PDF) of $C(\lambda, k)$ can be represented as follows:

$$f_{|C(\lambda,k)|^2}(x) = \frac{U(x)}{\sigma_B^2(\lambda, k) + \sigma_A^2(\lambda, k)} e^{-x/\left(\sigma_B^2(\lambda,k) + \sigma_A^2(\lambda,k)\right)} \tag{3}$$

where $\sigma_A^2(\lambda, k) = E|\ A(\lambda, k)\ |^2$ and $\sigma_B^2(\lambda, k) = E|\ B(\lambda, k)\ |^2$ are the PSDs of the clean speech signal and noise model respectively. The term *U(x)* defines the unit step function. The first order recursive equation is used to compute the smooth power spectrum of degraded speech data. It can be written as follows:

$$\mathrm{Z}(\lambda, k) = \beta \mathrm{Z}(\lambda - 1, \mathrm{k}) + (1 - \beta)\ |\ \mathrm{C}(\lambda, \mathrm{k})\ |^2 \tag{4}$$

Tracking the minima of degraded speech data was explained by Martin in [3]. The term β in (4) can be computed as

$$\beta = (T_{\mathrm{SM}} f_s / S - 1)\ /\ (T_{\mathrm{SM}} f_s / S + 1) \tag{5}$$

where T_{SM} is the window length of 0.2 seconds and S=128. By taking the value of sampling frequency, f_s of 8 kHz, S and T_{SM}, the value of β is computed. The tracking method was mainly depends on length of the minimum search window. A nonlinear method [6] is used

in the proposed method for tracking the minimum of degraded speech data by continuously taking the past spectral average values.

$$H_d(\omega) = \begin{cases} Z_{\min}(\lambda,k) = \gamma Z_{\min}(\lambda-1,k) + \frac{1-\gamma}{1-\xi}(Z(\lambda,k) - \xi Z(\lambda-1,k)); \\ \qquad\qquad \text{If } Z_{\min}(\lambda-1,k) < Z(\lambda,k) \\ Z(\lambda,k); \quad \text{Otherwise} \end{cases} \tag{6}$$

where $Z_{\min}(\lambda,k)$ is the local minima of the corrupted speech data power spectrum and γ and ξ constants which are considered while experimentation (γ = 0.999 and ξ = 0.8). The factor ξ used to control the adaptation duration of minima. To find the activity of speech in each frequency bin, the ratio of degraded speech spectrum to its local minima is considered and can be written as follows:

$$T_r(\lambda,k) = \frac{Z(\lambda,k)}{Z_{\min}(\lambda,k)} \tag{7}$$

A Bayes minimum-cost decision rule related to speech activity detection is given by

$$\frac{p(T_r \mid H_1)}{p(T_r \mid H_0)} \lessgtr_{H_0'}^{H_1'} \frac{c_{10} P(H_0)}{C_{01} P(H_1)} \tag{8}$$

where the priori probabilities are represented by $P(H_0)$ and $P(H_1)$ for speech absence and presence respectively. The cost for computing H_i' and H_j' is represented by the term c_{ij}. The probability of likelihood ratio $\frac{p(T_r|H_1)}{p(T_r|H_0)}$ is a monotonic function, therefore, the decision of (8) can be generalized as follows:

$$T_r(\lambda,k) \lessgtr_{H_0'}^{H_1'} \theta(k) \tag{9}$$

The value of $T_r(\lambda,k)$ is compared with threshold value. If the value of $T_r(\lambda,k)$ is greater than threshold, there is speech activity in that particular frequency bin else there is an absence of speech activity. This is mainly depending on the rule that the corrupted speech power spectrum is nearly equal to its local minima value when there is no speech activity. The speech activity decision making is described as follows:

$$D(\lambda,k) = \begin{cases} 1; & \text{Presence of speech if } \mathrm{Tr}(\lambda,k) > \theta(k) \\ 0; & \text{Otherwise} \end{cases} \tag{10}$$

where $\theta(k)$ is defined as frequency dependent threshold given experimentally. It can be approximated as

$$\theta(\mathrm{k}) = \begin{cases} 2 & 1 \le k \le LF \\ 2 & LF < k \le MF \\ 5 & MF < k \le f_s/2 \end{cases} \tag{11}$$

where MF and LF are the bin of frequencies with reference to *3kHz* and *1kHz* respectively and S is the sampling frequency. In [4], the value of $\theta(k)$ was fixed for all frequencies. From the above principle, the probability of speech presence $D(\lambda,k)$ is updated using first order recursion equation:

$$z(\lambda,k) = \mu_z z(\lambda-1,k) + (1-\mu_z) D(\lambda,k) \tag{12}$$

where μ_z is smoothing constant. The above recursion equation uses the correlation for presence of speech in adjacent frames or segments. By using the probability of speech presence approximation, the time-frequency dependent smoothing factor is calculated as:

$$\mu_{\mathrm{s}}(\lambda,k) \triangleq \mu_{\mathrm{d}} + (1-\mu_{\mathrm{d}})\, \mathrm{z}(\lambda,k) \tag{13}$$

where μ_d is a constant and the value of μ_d and μ_z are 0.85 and 0.25 respectively. The value of $\mu_s(\lambda, k)$ is always in the range of $\mu_d \leq \mu_s(\lambda, k) \leq 1$. The spectrum of noise estimate is updated as follows:

$$E(\lambda, k) = \mu_s(\lambda, k)\mathrm{E}(\lambda - 1, k) + (1 - \mu_s(\lambda, k)) \mid C(\lambda, k) \mid^2 \tag{14}$$

where $E(\lambda, k)$ is the approximation of noise power spectrum. The proposed algorithm is summarized as follows: Once the classification of presence of speech activity/non-speech activity is completed, the probability of speech presence/absence is updated. Using this probability, the time-frequency dependent smoothing factor is computed. At last, the noise spectrum approximation is updated using frequency-time dependent smoothing factor. The proposed noise estimator is used in speech enhancement technique to reconstruct the speech signal. Floored the negative values and ensured the conjugate symmetry for real reconstruction. The phase information is multiplied with whole frame of FFT and converted back to time domain using inverse fast Fourier transform. The above discussion of the proposed method is represented in terms of flowchart and pseudo code, which are shown in Fig. 1 and Algorithm 1 respectively.

Algorithm 1 A pseudo code of the proposed algorithm.

Input: Noisy speech signal: $c(n) \Leftarrow$ clean speech signal: $a(n)$ + noise model: $b(n)$.
Output: Spectrum of noise estimate and enhanced speech data: $E(\lambda, k)$ and $A(\lambda, k)$
1: // Formation of degraded speech signal: $c(n) = a(n) + b(n)$.
2: // FFT $\Rightarrow c(\lambda, k)$ and PDF $\Rightarrow f_{|C(\lambda,k)|^2}(x)$, of degraded speech signal by (2) and (3) respectively.
3: //Computation of smooth power spectrum of degraded speech data by first order recursive function:
$\mathrm{Z}(\lambda, \mathrm{k}) \Leftarrow \beta \mathrm{Z}(\lambda - 1, \mathrm{k}) + (1 - \beta) \mid \mathrm{C}(\lambda, \mathrm{k}) \mid^2$
4: // Computation of β by (5).
5: // Tracking the minima of degraded speech data: $H_d\ (\omega)$
6: // Computation of speech activity in each frequency bin:
$T_r(\lambda, k) \Leftarrow \frac{Z(\lambda,k)}{Z_{\min}(\lambda,k)}$
7: // Bayes minimum-cost rule for the decision:
$\frac{p(T_r|H_1)}{p(T_r|H_0)} \underset{H_0'}{\overset{H_1'}{\gtrless}} \frac{c_{10}P(H_0)}{C_{01}P(H_1)}$
$T_r(\lambda, k) \underset{H_0'}{\overset{H_1'}{\lessgtr}} \theta(k)$
8: //The speech activity decision making, $D(\lambda, k)$ and threshold, $\theta(k)$ are obtained using (10) and (11).
9: //$D(\lambda, k)$ is updated using first order recursion:
$z(\lambda, k) \Leftarrow \mu_z z(\lambda - 1, k) + (1 - \mu_z)\, D(\lambda, k)$
10: // Spectrum of noise estimate using the time-frequency dependent smoothing factor (13):
$E(\lambda, k) \Leftarrow \mu_s(\lambda, k)\mathrm{E}(\lambda - 1, k) + (1 - \mu_s(\lambda, k)) \mid C(\lambda, k) \mid^2$
11: Integration of computed spectrum of noise estimate with speech enhancement algorithm:
$A(\lambda, k) \Leftarrow \max\left\{\|C(\lambda, k)\|^2 - E(\lambda, k), vE(\lambda, k)\right\}$

4 Experimental results and analysis

Initially, the two performance metrics for the assessment of speech quality and intelligibility are discussed. Later, the integration procedure of the proposed noise estimation technique with a speech enhancement algorithm is explained. Further, the performance evaluation of the proposed method with existing noise estimation techniques in terms of speech quality and intelligibility after speech enhancement is demonstrated.

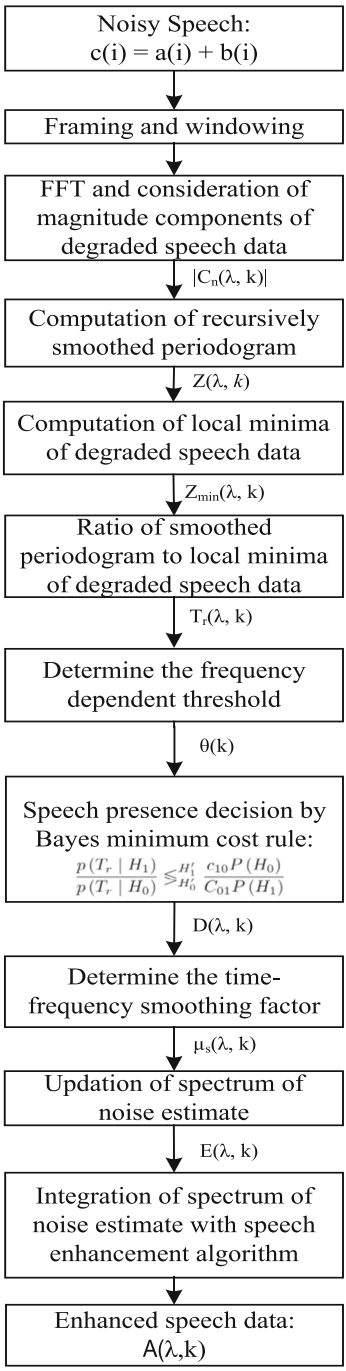

Fig. 1 Flowchart of the proposed algorithm

4.1 Performance metrics

The evaluation of proposed and existing techniques in terms of speech quality and intelligibility involves the consideration of two performance measures, namely segmental SNR (SSNR) and normalized covariance metric (NCM) [12, 14–17]. The ITU-T has established the SSNR as an objective metric for assessing speech quality. The NCM metric is formally characterized as the statistical measure of the covariance among the input and output response envelope signals [13]. The NCM is determined by subjecting the stimuli to a band pass filter that segments the signal into K bands across its bandwidth. The utilization of the Hilbert transform is employed for the computation of the envelope of every band, which is subsequently subjected to down sampling at a rate of 25kHz. Hence, the range of envelope modulation frequencies is restricted to 0-12.5kHz. Let us assume that $x_i(t)$ represents the down-sampled envelope in the i^{th} frequency band of the unaffected signal, while $y_i(t)$ denotes the down-sampled envelope of the transformed signal. The NCM of i^{th} frequency band can be written as follows:

$$r_i = \frac{\sum_t (x_i(t) - \mu_i)(y_i(t) - \nu_i)}{\sqrt{\sum_t (x_i(t) - \mu_i)^2}\sqrt{\sum_t (y_i(t) - \nu_i)^2}} \tag{15}$$

μ_{i} and $\mathrm{v_i}$ are the average values of clean and processed signals respectively. The SNR in each band is written as follows:

$$\mathrm{SNR}_i = 10\log_{10}\left(\frac{r_i^2}{1 - r_i^2}\right) \tag{16}$$

The transmission indices (TI) for i^{th} bands are calculated using the below equation:

$$\mathrm{TI}_i = \frac{\mathrm{SNR}_i + 15}{30} \tag{17}$$

The NCM index is generated by averaging the TI values across all frequency bands:

$$\mathrm{NCM} = \frac{\sum_{i=1}^{K} W_i \times \mathrm{TI}_i}{\sum_{i=1}^{K} W_i} \tag{18}$$

where W_i are the weights that are applied to each of the K bands. The proposed OSMC noise estimation technique is integrated with Wiener speech enhancement [16] technique using the following gain function:

$$G(\lambda, k) = \frac{A(\lambda, k)}{A(\lambda, k) + E(\lambda, k)\mu_k} \tag{19}$$

$$A(\lambda, k) = \max\left\{\|C(\lambda, k)\|^2 - E(\lambda, k), \upsilon E(\lambda, k)\right\} \tag{20}$$

where v = 0.001. The value of μ_k in (19) computed by the posterior SSNR [16].

4.2 A comparative analysis of proposed noise estimation technique with existing algorithms

The NOIZEUS [18] and Kannada speech databases are considered for the experimentation. The clean speech data is degraded by various types of noises, viz., babble, restaurant, and street noises at 5 dB and 10 dB SNR levels. The Kannada speech sentences are recorded in a controlled/uncontrolled environment and subsequently subjected to degradation by the same noises and same SNR level. For the analysis perspective, the proposed noise estimation

technique is compared with existing algorithms as follows: The main drawback of [3–5] is that, the noise estimation incurs a delay in computation and given poor reduction in noise for the speech data degraded by car, babble, street and train noises. The algorithms in [6, 7] failed to contrast between the increase in noise floor and speech power. The algorithms in [8–10, 19–21, 24] failed to update the noise estimation when the floor of noise increases suddenly and stays at a particular level. In comparison with the existing signal processing-based noise estimation techniques [3–10, 19–21, 24], the spectrum of estimation of noise using proposed method remains unaffected by a small search window. The proposed method

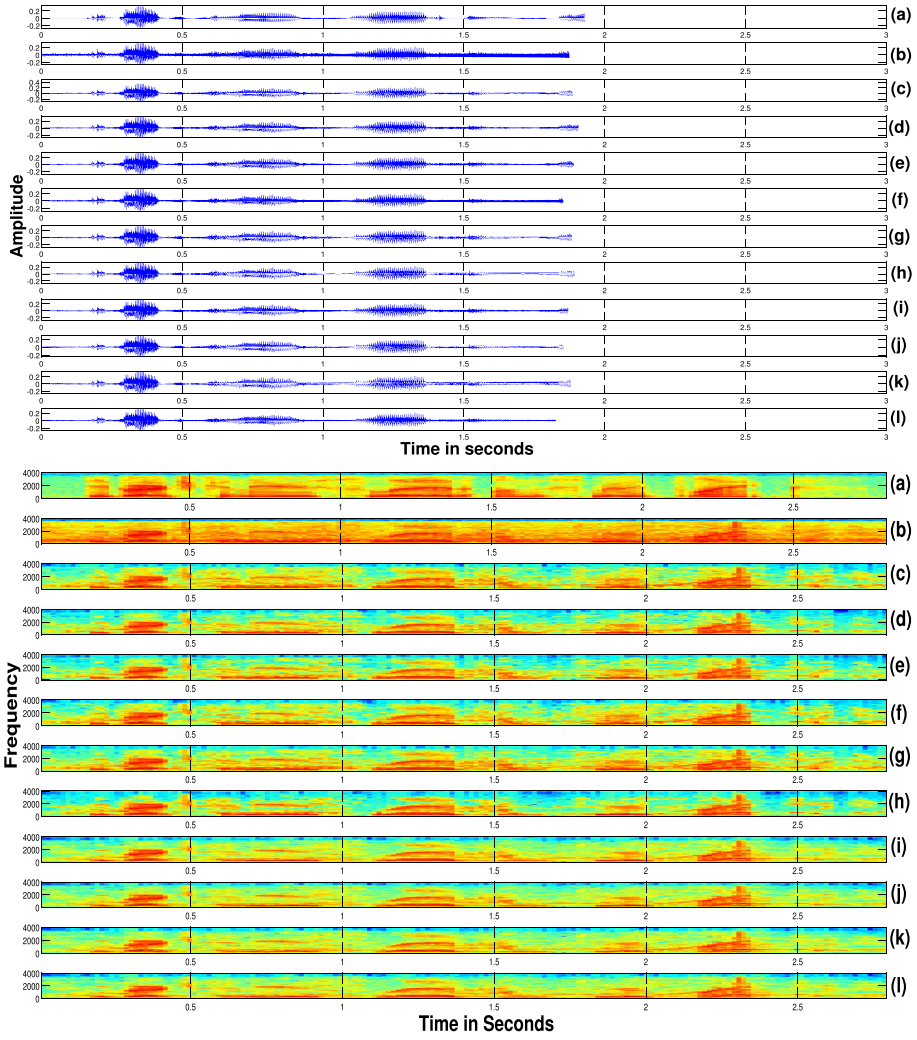

Fig. 2 The representation of input speech sentence from NOIZEUS database, output waveforms and spectrograms of: (a). Clean speech, (b). Noisy speech (babble noise at 5 dB SNR), Enhanced speech signal using (c). OSMS [3], (d). MCRA [4], (e). IMCRA [5], (f). Spectral minima tracking [6], (g). Weighted spectral averaging [7], (h). Connected time-frequency [8], (i). SS-VAD and MMSE-SPZC [9], (j). SS-VAD [19], (k). MSNE [20] and (l). Proposed algorithm (OSMC).

Table 1 Performance assessment of the proposed and existing methods in terms of SSNR for NOIZEUS and Kannada speech databases

Database	Noise Estimation Algorithm	SNR = 10 dB				SNR = 5 dB			
		Babble	Restaurant	Street	Average	Babble	Restaurant	Street	Average
NOIZEUS	OSMS [3]	2.4789	3.0603	2.8418	2.7937	−1.0001	−0.3390	0.0316	−0.4358
	MCRA [4]	3.2829	4.2015	3.8615	3.7819	−0.4216	0.3448	0.7626	0.2286
	IMCRA [5]	2.8607	4.1347	3.7458	3.5804	−0.8354	0.0484	0.5816	−0.0684
	Spectral minima tracking [6]	2.8894	2.9889	3.3707	3.0830	−0.6095	0.0947	0.0944	−0.1401
	Weighted spectral averaging [7]	2.3811	4.0067	2.8977	3.0951	−1.4232	−0.3564	0.1647	−0.5383
	Connected time-frequency [8]	2.5156	3.5384	3.9066	3.3202	−0.3242	0.3042	0.9284	0.3028
	SS-VAD and MMSE-SPZC [9]	3.0123	4.2158	3.8145	3.6808	−0.3241	0.3032	**0.9295**	0.3028
	SS-VAD [19]	2.9912	3.9983	3.2145	3.4013	−0.3343	0.2980	0.8291	0.2642
	MSNE [20]	3.0002	4.2188	3.6198	3.6129	−0.3221	0.3012	0.9185	0.2992
	Proposed Method (OSMC)	**3.3132**	**4.3281**	**3.9388**	**3.8600**	**−0.2582**	**0.5865**	0.9265	**0.4182**
Kannada	OSMS [3]	2.2401	2.6741	4.4064	3.1068	−0.6203	−0.208	0.8644	0.0120
	MCRA [4]	2.5988	2.9257	3.6821	3.0688	−0.5245	0.3394	0.476	0.0969
	IMCRA [5]	2.1931	2.3088	2.6174	2.3731	−0.4222	0.5070	0.4981	0.1943
	Spectral minima tracking [6]	2.0728	2.3760	3.5925	2.6804	−0.3826	−0.1198	0.7179	0.0718
	Weighted spectral averaging [7]	2.0568	2.6752	**4.5587**	3.0969	−0.7067	−0.1481	1.3323	0.1591
	Connected time-frequency [8]	1.7659	1.8270	2.0308	1.8745	−0.8533	−0.5894	0.1645	−0.4260
	SS-VAD and MMSE-SPZC [9]	2.3200	2.4256	2.6214	2.4556	−0.6968	−0.1081	**1.3443**	0.1798
	SS-VAD [19]	2.2990	2.3934	2.5912	2.4278	−0.7169	−0.1211	1.2343	0.1321
	MSNE [20]	2.5123	2.9910	3.9812	3.1615	−0.5345	0.3294	0.4886	0.0945
	Proposed Method (OSMC)	**2.6186**	**3.0666**	4.4170	**3.3674**	**−0.1767**	**0.3945**	0.7698	**0.3292**

Bold font indicates better performance

Table 2 Performance evaluation of the proposed and existing methods in terms of NCM for NOIZEUS and Kannada speech databases

Database	Noise Estimation Algorithm	SNR = 10 dB				SNR = 5 dB			
		Babble	Restaurant	Street	Average	Babble	Restaurant	Street	Average
NOIZEUS	MCRA [4]	0.8313	0.8390	0.8092	0.8265	0.6885	0.6553	0.7062	0.6833
	OSMS [3]	0.8241	0.8334	0.8015	0.8196	0.6661	0.6461	0.6985	0.6702
	Spectral minima tracking [6]	0.8127	0.8203	0.7857	0.8062	0.6703	0.6428	0.6899	0.6676
	IMCRA [5]	0.8286	0.8162	**0.8317**	0.8255	0.6826	0.6525	0.6896	0.6749
	Connected time-frequency [8]	0.8145	0.8273	0.7906	0.8108	0.6709	0.6403	0.6882	0.6664
	Weighted spectral averaging [7]	0.8226	0.8366	0.8018	0.8203	0.6809	0.6506	0.6986	0.6767
	SS-VAD [19]	0.8222	0.8312	0.7879	0.8137	0.6802	0.6501	0.6901	0.6734
	MSNE [20]	0.8212	0.8299	0.7832	0.8114	0.6879	0.6555	0.6891	0.6775
	SS-VAD and MMSE-SPZC [9]	0.8312	0.8345	0.7889	0.8182	0.6880	0.6489	0.6899	0.6756
	Proposed Method (OSMC)	**0.8431**	**0.8394**	0.8199	**0.8342**	**0.6892**	**0.6566**	**0.7108**	**0.6855**
Kannada	MCRA [4]	0.8183	0.8042	0.9181	0.8468	0.6954	0.7118	0.7851	0.7307
	OSMS [3]	0.8095	0.8016	0.8955	0.8355	0.6923	0.6846	0.7836	0.7201
	Spectral minima tracking [6]	0.7549	0.7765	0.8626	0.7980	0.6668	0.6645	0.7578	0.6963
	IMCRA [5]	0.8216	0.8138	0.9105	0.8486	0.7025	**0.7120**	0.7845	**0.7330**
	Connected time-frequency [8]	0.7467	0.7439	0.7840	0.7582	0.6468	0.6334	0.7277	0.6693
	Weighted spectral averaging [7]	0.8224	0.8038	0.9219	0.8493	0.6969	0.7006	0.7818	0.7264
	SS-VAD [19]	0.8214	0.8031	0.9121	0.8455	0.6968	0.6900	0.7801	0.7223
	MSNE [20]	0.8161	0.7867	0.8876	0.8301	0.6969	0.6944	0.7823	0.7245
	SS-VAD and MMSE-SPZC [9]	0.8212	0.8041	0.9211	0.8488	**0.6999**	0.6888	0.7850	0.7247
	Proposed Method (OSMC)	**0.8294**	**0.8143**	**0.9323**	**0.8586**	0.6925	0.6936	**0.7856**	0.7239

Bold font indicates better performance

uses a time-frequency dependent threshold value to compute the probability of speech presence/absence. In addition to this, the proposed algorithm works better for the floor of noise increases instantaneously and stays at a level.

The schematic representation of input-output waveforms and corresponding spectrograms of the proposed technique and existing algorithms are shown in Fig. 2. From the Figure, it can be inferred that the output waveform and spectrogram of the proposed method revealed that there is a significant amount of suppression of noise in the enhanced speech data compared to the competing algorithms. The Tables 1 and 2 show the performance evaluation of the proposed noise estimation technique with existing algorithms in terms of SSNR and NCM for NOIZEUS and Kannada speech databases respectively. From the Tables, it can be observed that, the proposed noise reduction algorithm outperformed the competing techniques in terms of the average values of SSNR and NCM. It is also observed that, there is a consistent and better improvement in the quality and intelligibility of speech data after enhancement across various types of noises at 5 dB and 10 dB SNR levels.

5 Conclusions

In this work, an algorithm was proposed for degraded speech enhancement by OSMC through iterative averaging under highly non-stationary noisy conditions. Unlike the state-of-the-art signal processing-based noise estimation techniques, the proposed algorithm updates the noise continuously under highly non-stationary noisy scenarios without using VAD. The experiments were conducted on NOIZEUS and Kannada speech databases for different types of noises and SNR levels. The SSNR and NCM metrics were used for the performance evaluation of proposed and existing algorithms for the assessment of speech quality and intelligibility. The evaluated results revealed that the proposed algorithm has given a consistent and better improvements over the existing algorithms in terms of SSNR and NCM values for different types of noises and SNR levels. In the future, it is interested to see whether the proposed technique could lead to better speech quality and intelligibility under negative SNR and deep noisy conditions. The future scope of the current work is to integrate the proposed noise elimination algorithm with an end-to-end ASR system to improve the accuracy of speech recognition under real-time conditions.

Funding Open access funding provided by Manipal Academy of Higher Education, Manipal.

Data Availability The manuscript has no associated data.

Declarations

Conflicts of interest The authors have no conflict of interests on the manuscript.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Rabiner L, Juang BH (1993) Fundamentals of speech recognition. Prentice-Hall Inc, New Jersey
2. Ramirez J, Gorriz JM, Segura JC (2007) Voice activity detection: Fundamentals and speech recognition system robustness. I-Tech Education and Publishing. https://doi.org/10.5772/4740
3. Martin Rainer (2001) Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans Speech Audio Process 9:504–512. https://doi.org/10.1109/89.928915
4. Cohen Israel (2002) Noise estimation by minima controlled recursive averaging for robust speech enhancement. IEEE Signal Process Lett 9:12–15. https://doi.org/10.1109/97.988717
5. Cohen Israel (2003) Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. IEEE Trans Speech Audio Process 11:466–475. https://doi.org/10.1109/TSA.2003.811544
6. Doblinger G (1995) Computationally efficient speech enhancement by spectral minima tracking in subbands. Citeseer 2:1513–1516. https://doi.org/10.21437/Eurospeech.1995-370
7. Hirsch H, Ehrlicher C (1995) Noise estimation techniques for robust speech recognition. In 1995 International conference on acoustics, speech, and signal processing, speech, signal processing vol. 1 pp. 153-156. https://doi.org/10.1109/ICASSP.1995.479387
8. Sorensen K, Andersen S (2005) Speech enhancement with natural sounding residual noise based on connected time-frequency speech presence regions. EURASIP J Adv Signal Process 18:2954–2964. https://doi.org/10.1155/ASP.2005.2954
9. Thimmaraja Yadava G, Jayanna HS (2018) Speech enhancement by combining spectral subtraction and minimum mean square error-spectrum power estimator based on zero crossing. Int J Speech Technol (IJST) Springer 22:639–648. https://doi.org/10.1007/s10772-018-9506-9
10. Thimmaraja Yadava G, Jayanna HS (2020) Enhancements in automatic Kannada speech recognition system by background noise elimination and alternate acoustic modelling. Int J Speech Technol (IJST) Springer 23:149–167. https://doi.org/10.1007/s10772-020-09671-5
11. Thimmaraja Yadava G, Jayanna HS (2017) A spoken query system for the agricultural commodity prices and weather information access in Kannada language. Int J Speech Technol (IJST) Springer 20:635–644. https://doi.org/10.1007/s10772-017-9428-y
12. Kates James M, Arehart Kathryn H (2005) Coherence and the speech intelligibility index. J Acoust Soc Am 117:2224. https://doi.org/10.1121/1.1862575
13. Ma J, Hu Y, Loizou PC (2009) Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J Acoust Soc Am 125:3387–3405. https://doi.org/10.1121/1.3097493
14. John H. L. Hansen, Bryan L. Pellom (1998) An effective quality evaluation protocol for speech enhancement algorithms. Proceedings 5th international conference on spoken language processing (ICSLP 1998) Sydney, Australia. https://doi.org/10.21437/ICSLP.1998-350
15. Stahl V, Fischer A, Bippus R (2000) Quantile based noise estimation for spectral subtraction and Wiener filtering. In 2000 IEEE international conference on acoustics, speech, and signal processing. Proceedings vol 3 pp 1873–1875. https://doi.org/10.1109/ICASSP.2000.862122
16. Yi Hu, Loizou PC (2004) Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE Trans Speech Audio Process 12:59–67. https://doi.org/10.1109/TSA.2003.819949
17. Hu Y, Loizou PC (2008) Evaluation of objective quality measures for speech enhancement. IEEE Trans Audio Speech Lang Process 16:229–238. https://doi.org/10.1109/TASL.2007.911054
18. Hu Y, Loizou PC (2007) Subjective evaluation and comparison of speech enhancement algorithms. Speech Commun 49:588–601. https://doi.org/10.1016/j.specom.2006.12.006
19. Thimmaraja Yadava G, Nagaraja BG, Jayanna HS (2021) Speech enhancement and encoding by combining SS-VAD and LPC. Int J Speech Technol Springer 24:165–172. https://doi.org/10.1007/s10772-020-09786-9
20. Tan Zheng-Hua, Sarkar Achintya K R, Dehak Najim (2020) rVAD: An unsupervised segment-based robust voice activity detection method. Comput Speech Lang 59:1–21. https://doi.org/10.1016/j.csl.2019.06.005
21. Jaiswal RK, Yeduri SR, Cenkeramaddi LR (2022) Single-channel speech enhancement using implicit Wiener filter for high-quality speech communication. Int J Speech Technol Springer 25:745–758. https://doi.org/10.1007/s10772-022-09987-4
22. Bahrami M, Faraji N (2021) Minimum mean square error estimator for speech enhancement in additive noise assuming Weibull speech priors and speech presence uncertainty. Int J Speech Technol Springer 24:97–108. https://doi.org/10.1007/s10772-020-09767-y
23. Roy S, Paliwal KK (2021) A noise PSD estimation algorithm using derivative-based high-pass filter in non-stationary noise conditions. EURASIP J Audio Speech Music Process 32. https://doi.org/10.1186/s13636-021-00220-9

24. Gupta M, Singh RK, Singh S (2023) Analysis of optimized spectral subtraction method for single channel speech enhancement. Wireless Pers Commun 128:2203–2215. https://doi.org/10.1007/s11277-022-10039-y
25. Ghorpade K, Khaparde A (2023) Single-channel speech enhancement using single dimension change accelerated particle swarm optimization for subspace partitioning. Circuits Syst Signal Process 42:4343–4361. https://doi.org/10.1007/s00034-023-02324-3
26. Liang Ruiyu, Xie Yue, Cheng Jiaming, Tang Guichen, Sun Shinuo (2021) Real-time speech enhancement algorithm for transient noise suppression. Multimed Tools Appl 80:3681–3702. https://doi.org/10.1007/s11042-020-09849-8
27. Thimmaraja Yadava G, Nagaraja BG, Jayanna HS (2022) A spatial procedure to spectral subtraction for speech enhancement. Multimed Tools Appl 81:23633–2364. https://doi.org/10.1007/s11042-022-12152-3
28. Thimmaraja Yadava G, Nagaraja BG, Jayanna HS (2023) Amalgamation of noise elimination and TDNN acoustic modelling techniques for the advancements in continuous Kannada ASR system. Multimed Tools Appl. https://doi.org/10.1007/s11042-023-16100-7
29. Jainar SJ, Sale PL, Nagaraja BG (2020) VAD, feature extraction and modelling techniques for speaker recognition: a review. Int J Signal Imaging Syst Eng 12:1–18. https://doi.org/10.1504/IJSISE.2020.113552
30. Nagaraja BG, Jayanna HS (2016) Feature extraction and modelling techniques for multilingual speaker recognition: a review. Int J Signal Imaging Syst Eng 9:67–78. https://doi.org/10.1504/IJSISE.2016.075000
31. Nagaraja BG, Jayanna HS (2013) Multilingual speaker identification by combining evidence from lpr and multitaper mfcc. J Intell Syst 22:241–251. https://doi.org/10.1515/jisys-2013-0038

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Raghudathesh G P[1] · Chandrakala C B[2] · Dinesh Rao B[3] · Thimmaraja Yadava G[4]

Raghudathesh G P
raghudathesh.gp@manipal.edu

Dinesh Rao B
dinesh.rao@manipal.edu

Thimmaraja Yadava G
thimrajyadav@gmail.com

1 Manipal School of Information Sciences, Manipal Academy of Higher Education, Manipal 576104, Karnataka, India

2 Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, Karnataka, India

3 Manipal School of Information Science, Manipal Academy of Higher Education, Manipal 576104, Karnataka, India

4 Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bengaluru 560064, Karnataka, India