Analysis Modification synthesis based Optimized Modulation Spectral Subtraction for speech enhancement

Pavan D. Paikrao[1*], Sanjay L. Nalbalwar[2],

Abstract—Traditional analysis modification synthesis (AMS) is fairly applied for spectral subtraction along with Short Time Fourier Transform. Based on this AMS method, we proposed an approach for modified modulation spectral subtraction. Results reported in previous studies shows that the modulation spectral subtraction performs better for speech courted by additive white Gaussian noise to improve speech quality. It gives improved speech quality scores in stationary noise, but it fails to give improved speech quality in the real time noise environment. Also, the computational cost of existing modulation domain spectral subtraction methods is high. Thus we propose an approach of applying minimum statistics noise estimation technique on the real modulation magnitude spectrum along with optimized noise suppression factor and spectral floor to improve speech quality in the real time noise environment. Finally, the objective, subjective and intelligibility evaluation metrics of speech enhancement indicates that the proposed method achieves better performance than the existing spectral subtraction algorithms across different input SNR and noise type along with improved computational time. Computation time is improved by 57.13% as compared to traditional modulation domain spectral subtraction method. The modulation frame duration of 128 ms is found to be a good compromise between shorter and longer frame duration, which gives improved results.

***Keywords*—**Optimized modulation spectral subtraction, speech enhancement, Analysis modification synthesis, Noise.

I. INTRODUCTION

The use of speech enhancement has a spurred great interest in many fields such as speech recognition, feature extraction, hearing aid devices, etc. Human exhibits great capability to differentiate various sounds in noisy environments. But, unfortunately performance of these speech enhancement systems decays when speech is corrupted with stationary or non-stationary background distortions. Speech enhancement is nothing but a process of improving the quality of noisy speech. It means a speech enhancement system reduces that additive noise which corrupts the original speech and makes it annoying to the listener. Thus, in noisy environment conditions there is a crucial need to improve the performance of these systems.

Several researchers have proposed different classical speech enhancement techniques [1,2,3,4,5] which remove additive noise.

[1] Pavan D. Paikrao is with Department of Electronics & Tele Comm. Engg., Dr. Babasaheb Ambedkar Technological University, Lonere, Dist. Raigad, MS, India. (Corresponding author e-mail: pavan242batu@gmail.com
[2]Sanjay L. Nalbalwar is with Department of Electronics & Tele Comm. Engg., Dr. BabasahebAmbedkar Technological University, Lonere, Dist. Raigad, MS, India.

The generalized approach for speech enhancement algorithm is to modify or enhance spectral component and reduce background noise. The spectral subtraction method proposed by Berouti [1] and [2] is classical noise suppression methods. These methods use a spectral floor threshold and noise suppression factor which governs the amount of over subtraction in accordance with the SNR level of the input noisy signal. It reported different values of noise suppression factors so as to have different efficient noise suppression paradigm. It is the subject of research to adjust these parameters in different noisy environmental conditions for enhanced speech quality.

Over last few decenniums, many speech enhancement methods have been investigated that includes time and frequency domain modifications. According to Kamath's Multi Band Spectral Subtraction (MBSS) [6], the speech signal is not affected uniformly by additive noise over the entire spectrum. Low frequency components which contain most of the speech signal energy get affected more easily than high frequency components by noise. In this method, the speech signal is divided into a number of non-overlapping bands and spectral subtraction is carried out independently in each band for speech enhancement.

More recently, a phase-aware multi-band complex spectral subtraction (MBCSS) method introduced by [7], deals with single channel speech enhancement by improved phase at low input SNR. MBCSS computes spectral amplitude of clean speech signal using phase of clean and noisy speech signals and uses the estimated phase of the clean speech signal for signal reconstruction in the time domain. MBCSS method can dynamically adapt itself according to the varying levels of non-stationary noise and the phase components of speech. Noise is separated by a single channel source separation technique based on group-delay deviation which is effectively utilized in the spectral subtraction method.

Many single channel speech enhancement methods employ analysis, modification synthesis (AMS) technique [8,9,10,11]. AMS framework is applied in acoustic domain spectral subtraction to reduce additive noise. Here, we are dealing with the enhancement of speech corrupted by additive noise. In speech enhancement process, this additive noise can be put into two categories as stationary noise, i.e. additive white Gaussian noise (AWGN) and non-stationary noise (real time background noise). AWGN is linear and Time Invariant. While real time background noise is produced by dynamic environments. For example car noise, train noise, airport noise, or many other man made noise, etc. are non-stationary noises. In a non-stationary environment, noise estimation is a difficult task if the noise power

changes during voice presence. Stationary noise on the other hand can be easily evaluated mathematically and can be reduced to the greatest extent by proper design of speech enhancement system. The single channel speech enhancement modulation spectral subtraction (ModSpecSub) method [11] reported improved speech quality, especially in AWGN noise along with reduced background noise. ModSpecSub employs Voice activity detection (VAD) algorithm to estimate noise using recursive averaging of non-speech frames, which is applied in generalized spectral subtraction thus it is computationally expensive. ModSpecSub technique gives improved objective scores in AWGN but in the real time (non-stationary) background noise environment, objective scores found to be reduced. The audio stimuli generated by ModSpecSub method gives reduced background noise and musical artifact, however speech slurring is observed during listening tests.

In this paper, we focus on the enhancement of single channel speech corrupted by real time background noise environment and to reduce computational time in modulation domain spectral subtraction. Thus, we introduce an approach of applying the minimum statistics noise estimation method in modulation domain. As a result, we achieve reduced speech slurring, improved speech quality and reduced computational time. We employ analysis modification synthesis framework in which after computing Short Time Fourier Transform (STFT), the complex spectrum is generated. Now this spectrum is bifurcated in the real and imaginary spectrum and the only real spectrum is further processed discarding the imaginary spectrum (in both acoustic and modulation domain processing). Thus the proposed approach exhibits lower computational time than the computational time of ModSpecSub [11] method. The proposed algorithm is optimized in terms of modulation frame duration and several parameters for improved speech quality. The minimum statistics noise estimation method is incorporated with proposed optimized modulation spectral subtraction (OMSS). The proposed algorithm is evaluated using NOIZEUS [12] speech corpus, which is a database of different noisy signal conditions at different input SNR and is freely available. Furthermore, we have performed both subjective and objective evaluation of proposed OMSS method that proves consistent speech quality improvements at various input SNRs.

II. Analysis-modification-synthesis (AMS)

A. AMS Framework

Analysis modification synthesis (AMS) method [8,9] is an efficient method for signal enhancement. AMS uses following steps. First, framing of the input speech signal with suitable window function and Second, STFT of widowed frames with some frame shift. Third, inverse Fourier Transform and fourth retrieving signal by overlap and add (OLA) method [10]. Let's consider our speech is as follows

$$x(n) = s(n) + N(n) \tag{1}$$

x(n),s(n) and N(n) are input sampled noisy speech signal, pure speech, and disturbing noise signal respectively. Whereas n is the discrete time index. Since speech signal is non-stationary in nature.

In an AMS framework, speech is processed over a short frame duration by using STFT [8,9]. Now from the definition of STFT, spectrum of noise corrupted speech is

$$X(n,k) = \sum_{l=-\infty}^{+\infty} x(n)w(n-l)e^{-j2\pi kl/M} \tag{2}$$

Where l is an acoustic frame number, k is an index of discrete acoustic frequency, M is acoustic frame duration in samples and w (n) is an analysis window function. We applied modified Hanning window [8] at both acoustic and modulation domains which is found to be efficient as compared to other window function.

The AMS framework is repeated after acoustic domain processing to work in modulation domain. Thus we tried to apply spectral subtraction in modulation domain [11] speech signal with the speech enhancement technique like [1,2] as shown in Fig. 1. Thus Eq.1 can be represented by applying STFT, as

$$X(n,k) = S(n,k) + N(n,k) \tag{3}$$

Where X(n, k), S(n, k) and N(n, k) are spectrum of input noisy speech, pure speech, and disturbing noise respectively. In general, these transforms can also be represented as acoustic magnitude spectrum and acoustic phase spectrum as

$$X(n,k) = |X(n,k)|e^{j<X(n,k)} \tag{4}$$

Where |X (n, k) | indicates an acoustic magnitude spectrum and <X (n, k) indicates an acoustic phase spectrum. The STFT algorithm is computationally efficient and can be implemented for real-time application. After framing the signal by using an appropriate windowing technique, the spectral modification is applied to STFT magnitude spectrum.

B. Conventional Spectral subtraction

Most of the Spectral subtraction approach estimates enhanced speech by subtracting short time spectral amplitude of the estimated noise from disturbing noise signal. This subtraction may give negative values depending on magnitudes of current frame noise spectra and estimated disturbing noise spectra. To avoid this inconsistency the noise flooring as a function of the over-subtraction factor is employed. The enhance spectrum is

$$\hat{S}(n,k) = \left(X(n,k)^{(\Upsilon)}\right) - \alpha\left(N(n,k)^{(\Upsilon)}\right) \tag{5}$$

Noise floor B_N is estimated as follows

$$B_N = (\beta(N(n,k)^{(\Upsilon)}))^{(1/\Upsilon)} \tag{6}$$

Where α and β are over-subtraction factor and noise floor factor respectively. N(*n,k*) is noise estimate and γ is spectral subtraction domain. For γ=1, it is magnitude, spectral subtraction and γ=2, it is a power spectral subtraction. The enhanced estimated of clean speech S(*n,k*) given by Berouti [1] is

$$\mathrm{S}(n,k)=\max\{\hat{\mathrm{S}}(n,k),\mathrm{B}_N(n,k)\} \tag{7}$$

C. Conventional noise estimation

Most of the speech enhancement methods use the VAD algorithm. VAD algorithm is used to detect whether the input signal is speech or noise only. That means VAD categories every frame in 1 (speech presence) or 0 (speech absence).

ModSpecSub[11] method obtains noise estimate by averaging over initial silence frames. Now the time average noise spectrum can be obtained from the frames when a speech frame is absent i.e. only noise is present. This estimated noise we termed as noise estimation over an initial silence frame. Let's consider the speech sample stimuli sp02 of NOIZEUS speech corpus [12], which is of total duration 3 s and the initial silence period is 0.7 s. Thus, these initial silence frames over 0.7 s duration is used for noise estimation.

$$|N(n,k)|^{\Upsilon}=\frac{1}{k}\sum_{i=0}^{k-1}|Xi(n,k)|^{\Upsilon} \tag{8}$$

Where $Xi(n,k)$ spectrum of ith is input initial silence frame. Here it is assumed that selected frames are noise only frame. Now this noise estimate is updated during speech absence, using the averaging rule of Virag [4]. ModSpecSub [11] used initial silence frame for pre-estimating noise. However, this is unrealistic situation. Initial silence is not present in real time background noise environment. Therefore the noise estimate with this method is not appropriate in real non stationary environment.

This process increases the computational load of the system. So to reduce this computational load we propose an approach to apply minimum statistic noise estimation [13,14,15] in modulation domain.

D. Overlap-add (OLA) method

As introduced by Griffin and Lim [8], to reconstruct the modified signal after inverse Fourier transform, OLA is applied in both acoustic and modulation domain synthesis processing. In this reconstruction step, the inverse DFT of each frame in discrete STFT is computed. This is then divided by analysis window. The intuition is to remove the mismatching between overlapped frames. Thus the OLA method can be expressed as

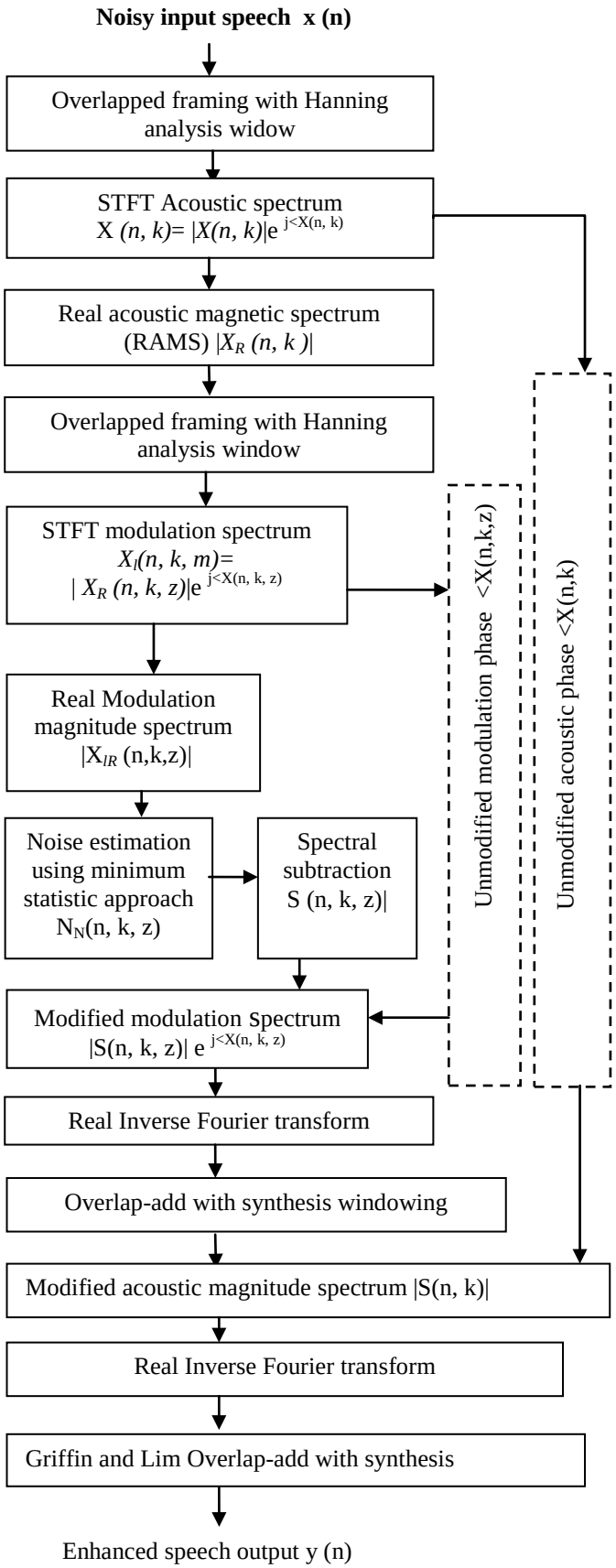

Fig. 1 *Flow chart of a proposed OMSS, AMS-based speech enhancement method.*

$$\frac{1}{w(0)}\sum_{p=-\infty}^{+\infty}\left[\sum_{k=0}^{N-1}X(p,k)e^{-j\frac{2\Pi kn}{N}}\right] \quad (9)$$

Where w(n) is a synthesis window function.

III. MODULATION DOMAIN PROCESSING

A. Method

The modulation spectrum is obtained from the traditional AMS based acoustic spectrum discussed in section 2.1. It is formulated from the each frequency domain transform achieved from acoustic spectrum transform using STFT. The each frequency component achieved in the acoustic domain transform is processed frame by frame using another AMS framework across time.

Now the modulation spectrum can be formulated as

$$X(n,k,z)=\sum_{l=-\infty}^{+\infty}x(n)w(n-l)e^{-j2\pi kl/N} \quad (10)$$

Where n is an acoustic frame number, k is the index of discrete acoustic frequency, z is termed as an index of the discrete modulation frequency. N is modulation frame duration, w(n) modulation analysis frame window function.

In modulation domain the STFT is computed at given acoustic frequency from time series of real acoustical spectral magnitudes $|X_R(n, k)|$ at that frequency. Hanning window with optimal frame duration of 128 ms and frame shift of 16 ms is used in modulation domain.

B. Modification

Appropriate noise estimate is an essential step in spectral subtraction. The effect of different noise estimation method on our modified modulation spectral subtraction is studied. Optimal noise estimates in speech enhancement so as to reduce computational complexity is needed. Extensive experimental evaluation based on noise estimation techniques in modulation domain spectral subtraction done. First, noise estimation using initial silence frame and second, minimum statistic noise estimation approach. The first approach employs a VAD algorithm to update the noise during non-speech frames and pause between utterances. Thus the computational load is greater. In proposing methods, experimental evaluation, it is observed that at large frame duration and frame shift, no considerable effect of noise updating in found in the modulation domain processing. Thus we avert the use of the VAD [17] algorithm for noise updating and apply minimum statistic noise estimation approach in the modulation domain to reduce the computational load on the proposed access.

The minimum statistic method of noise estimation gives improved speech quality.

In the proposed OMSS approach following steps are involved as shown in Fig. 1.

Step I: In the pre-emphasis step, noisy input speech signal (no mean subtracted) is segmented into overlapping acoustic frames using analysis window duration of 32 ms and STFT is applied to each frame which gives complex acoustic spectrum X (n, k).

This STFT of the speech signal is a complex valued spectrum build in with a real and imaginary part as shown in Eq.(11).

$$X(n,k)=X_R(n,k)+i.X_I(n,k) \quad (11)$$

Where $X_R(n,k)$ is real part and $X_I(n,k)$ is imaginary part of acoustic spectrum $X(n,k)$. Now the real part $X_R(n,k)$ of this complex acoustic spectrum is computed (discarding imaginary part) and we terms it as Real Acoustic Magnitude Spectrum (RAMS) denoted as $|X_R(n, k)|$.Where |. | denote absolute value of the complex number. Phase is also estimated from this RAMS, which will be combined later during the synthesis stage.

Step II: Now the RAMS is applied to the secondary AMS framework as described in section 2.1. The noisy envelope RAMS $|X_R(n, k)|$ is segmented into overlapped modulation frames with modulation frame duration of 128ms duration and second STFT is applied along the time axis (at each frequency) to form the complex spectrum X (n, k, z). It can be represented

$$X(n,k,z)=X_R(n,k,z)+i.X_I(n,k,z) \quad (12)$$

Where *z* is a modulation frame index and k is the acoustic frequency index. Now the real part of this complex modulation spectrum X(n, k, z) is computed, we term it as Real Modulation Magnitude Spectrum (RMMS) $|X_R(n, k, z)|$ by discarding imaginary part. The modulation domain phase is estimated from this RMMS which will be combined later during the synthesis stage.

In modulation domain spectral subtraction, large frame duration up to 280 ms can be applied. But at this longer frame duration stationarity needs to be assume (in contradictory to non-stationary nature of speech), which yields speech temporal slurring. Also, due to longer frame duration, the computational load increases. To minimize the temporal speech slurring and the computational load, optimal modulation frame duration was decided to 128 ms and frame shift of 16 ms in modulation domain processing by repeated experiments. It means for this modulation frame duration of 128 ms, an improved performance of several objective scores [17,19] such as Log Likelihood Ratio (LLR), Weighted Spectral Slope(WSS), SNRseg, Csig., Covl., as shown in Table I, Table II and Fig. 3, Fig. 4 is observed. The speech intelligibility score Short-Time objective intelligibility (STOI) in [19] also significantly improved as shown in Fig. 4.

Step III: The appropriate noise estimation is a crucial part of speech enhancement technique. In the conventional speech enhancement methods, noise estimate is obtained from the input noisy speech signal. In contrast to conventional way we applied RMMS frames for noise estimation. It means noise estimation from RMMS for the spectral subtraction in modulation domain is applied to the proposed approach. Here as shown in Fig. 2, we studied the effect of different noise estimation method, such as minimum statistics [13,14,15], Unbiased MMSE noise estimation [16] on proposed Optimized modulation spectral

subtraction (OMSS) method. Among these methods, noise estimation using RMMS spectrum by minimum statistical method is found to give improved speech quality and intelligibility.

At a later stage after modulation domain spectral subtraction, modulation domain phase is recombined with enhanced signal $S^{\wedge}(n,k,z)$ to form modified spectrum as shown in Fig. 1. The enhanced speech signal, Y(n) is constructed by taking the inverse STFT of the modified modulation spectrum followed by least-squares overlap-add synthesis.

Modulation domain spectral subtraction: For Spectral subtraction in modulation domain, we apply

$$\hat{S}(n,k,z)=\left(X(n,k,z)^{(\Upsilon)}\right)-\alpha\left(N(n,k,z)^{(\Upsilon)}\right) \tag{13}$$

Where $\hat{S}(n,k,z)$ is an estimate of the clean speech signal,$|XR(n,k,z)|$ is RMMS and $N(n,k,z)$ is the noise spectrum obtained using minimum statistics noise estimation algorithm. α is the over-subtraction factor which controls the amount of subtraction of noise estimate from the noisy speech signal. The over-subtraction factor α conventionally can be used between 0-6. For minimum statistics method [10,14,15] of noise estimation, this should be between 0 and 3. The optimized results were obtained at α=1. However α for unbiased MMSE noise estimator [12], is found to be optimized between 0-1. For α=0.1 gives improved objective scores, but for α=1, gives reduced objective scores.

We apply the over-subtraction factor $0.1 \leq \alpha \leq 3$. The following values were used in the implementation, α=1, β= 0.0001, γ=2. It is found that spectral subtraction gives optimized objective scores at γ=2, α=1 as shown in Fig. 3, 4, 5 and 6.

C. Noise estimation

Conventional noise estimation using initial silence frames of input noisy speech signal: The conventional ModSpecSub employ VAD [17] on the estimate of initial silence frames to update the noise estimate, which gives reduced speech quality scores in the non-stationary environment and computational load increases.

Noise estimation using the minimum statistics method: In this method [14] the power spectral density (PSD) of non-stationary, especially additive noise is estimated from the input noisy speech.

Reason: why the minimum statistic method in modulation domain ?: - In modulation domain processing the frame duration is large as compared to that in the acoustic domain. Thus, over this large frame duration in modulation domain, the use of VAD yields no effect on speech and non-speech frames which is applied in conventional ModSpecSub method [11] to update noise in non-speech frames.

In [13,14,15] the PSD of noise is estimated without using Voice activity detection. Instead, it tracks the spectral minima over each frame independent of speech and non-speech frames.

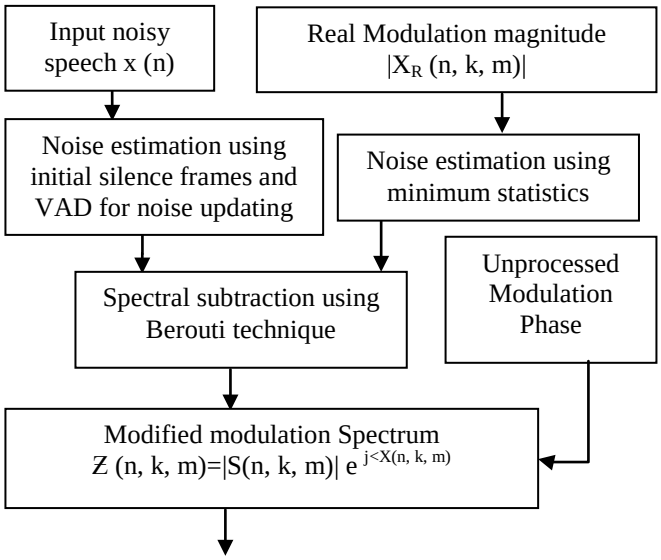

Fig. 2 *Noise estimation and spectral subtraction paradigm.*

Therefore, computational speed is also improved.

The smooth noise PSD is shown by

$$P(n,k)=\alpha P(n-1,k)+|XlR\ (n,k)|^2 \tag{14}$$

Where n is time index, k is frequency index (k € {0,1,.. L-1}), L in the modulation FFT index and α* is smoothing parameter. Here in this approach to minimize the error between estimated PSD,P(n, *k*) and true estimate $N^2(n,k)$ of noise, the conditional mean square error is estimated as follow.

$$E\{(P(n,k)-N^2(n,k))^2\,|\,(P(n-1,k)\} \tag{15}$$

Now putting $E\{X(n,k)^2\}=N^2(n,k)$ and $E\{X(n,k)^4\}=2N^4(n,k)$ It gives

$$E\{(P(n,k)-N^2(n,k))^2\,|\,(P(n-1,k)\}$$
$$=\alpha(n,k)(P(n,k)-N^2(n,k))^2$$
$$+N^4(n,k))^2(1-\alpha(n,k))^2 \tag{16}$$

Now the short term PSD is calculated as

$$P(n,k)=(1-\alpha^*)\sum_{i=0}^{\infty}\alpha^* i\,|XlR(n-i,k)|^2 \tag{17}$$

Now the minimum estimate of P(*n*,*k*) is termed as

$$B_{\min}^{-1}(n,k)=E\{P_{\min}(n,k)\}\big|_{N^2(n,k)=1} \tag{18}$$

This minimum function is written in terms of inverse normalized variance $q_{eq}(n,k)$ from [15, Sec. 7.2] as

$$B_{\min}^{-1}(n,k)\approx 1+(D-1)\frac{2}{\tilde{q_{eq}}(n,k)}\Gamma\left(1+\frac{2}{q_{eq}(n,k)}\right)^{h(D)} \tag{19}$$

Where D is the length of the minimum search window and $q_{eq}^{\sim}(n,k)$ is scaled version of $q_{eq}(n,k)$. Here $q_{eq}(n,k)=2$ for $B_{\min}=D$ is employed in Eq. (18). The constant approximation values are considered as $D=1$.The gamma function Γ(.) taken from [15]. Finally, the unbiased noise is derived as

$$N_N^{\wedge 2}(n,k)=\frac{P_{\min}(n,k)}{E\{P_{\min}(n,k)\}_{|N^2(n,k)=1}} \quad (20)$$

IV. EXPERIMENTAL EVALUATION RESULTS

A. Database used

In our experiments, we employ the NOIZEUS speech corpus database [12,17]. The basic premise of a database like NOIZEUS is to make recordings of more realistic noises at different input SNRs available to researchers. Speech corpus is composed of 30 IEEE phonetically-balance sentences of six speakers (3 male and 3 females). The speech sentences are sampled at 8 kHz. For our experiments, we used the corpus noisy stimuli of real time noise environment such as airport, babble, car, restaurant, station and train background noises at various input SNRs.

B. Experimental setup

We have used Intel core i3 processor in the 2.4 GHz clock frequency personal computer (PC). The proposed approach of spectral subtraction in modulation domain is implemented in MATLAB R2009.The input noisy speech signal is pre-emphasized. Many speech enhancement methods make the input signal zero mean, but we have only made our input signal in raw form and did not subtract the mean of the input signal from it. For simplified declaration, we termed acoustic domain as STFT of the input speech signal and modulation domain as STFT of time series of acoustical spectral magnitude at each frequency. In the acoustic domain processing input signal is segmented by using Hanning window of 32 ms with 40% overlap. Then each frame of noisy input is getting transformed into frequency domain with 256 point FFT.

C. Objective evaluation:

LLR and WSS are strongly co-related to the distortion in speech and weakly correlated with reduction in noise.

For the best performance, these objective scores should be low. Lowest LLR and WSS scores for proposed OMSS method show that the signal quality is improved. Further speech distortion is low. Table I and Table II shows the average (mean) results of the LLR and WSS scores for 30 IEEE sentences for different spectral subtraction methods like Paliwals method [11], Samui's MBCSS [7], Boll's method [2], Berouti's method [1], and Kamath's MBSS method [6] respectively.

Table I Results of mean LLR scores

		Spectral enhancement technique					
Noise Type	Input SNR (dB)	Proposed OMSS	Paliwal's ModSpecSub [11]	Samui's MBCSS [7]	Boll's [2]	Berouti's [1]	Kamath's MBSS [6]
Car	0	**1.04**	1.07	1.71	1.8	1.61	1.68
	5	**0.83**	0.91	1.44	1.18	1.41	1.26
Babble	0	**1.00**	1.05	1.74	1.61	1.69	1.61
	5	**0.76**	0.91	1.73	1.18	1.29	1.38
Airport	0	**0.96**	1.03	2.32	1.92	1.83	1.71
	5	**0.77**	0.91	2.39	1.36	1.48	1.27
Station	0	**0.99**	1.03	2.12	1.22	1.63	1.59
	5	**0.72**	0.80	1.39	1.32	1.78	1.21
Exhibition	0	1.36	1.42	1.17	1.82	1.73	1.21
	5	**0.99**	1.12	1.59	1.16	1.23	1.12

From Table I, for babble noise at 5dB input SNR 17.46% LLR improvement is reported as compared to ModSpecSub.

Table II Results of mean WSS scores

		Speech enhancement technique					
Noise Type	Input SNR (dB)	Proposed OMSS	Paliwals ModSpecSub [11]	MBCSS [7]	Boll's [2]	Berouti's [1]	Kamath's MBSS [6]
Car	0	60.17	57.71	56.17	73.21	75.34	58.32
noise	5	**48.56**	50.20	56.49	65.97	71.66	50.44
Babble	0	**68.21**	77.11	72.87	93.71	104.58	76.57
noise	5	**52.33**	63.96	61.94	70.58	73.99	54.3
Airport	0	**72.51**	90.98	63.92	80.17	100.41	63.92
noise	5	**59.18**	85.51	63.11	71.47	85.78	63.11
Station	0	**62.72**	63.47	70.93	90.77	102.54	74.92
noise	5	**51.27**	54.02	61.16	70.47	75.28	73.12
Exhibition	0	**64.50**	74.49	53.92	86.97	79.41	83.19
noise	5	**54.01**	65.87	59.57	71.47	75.71	62.11

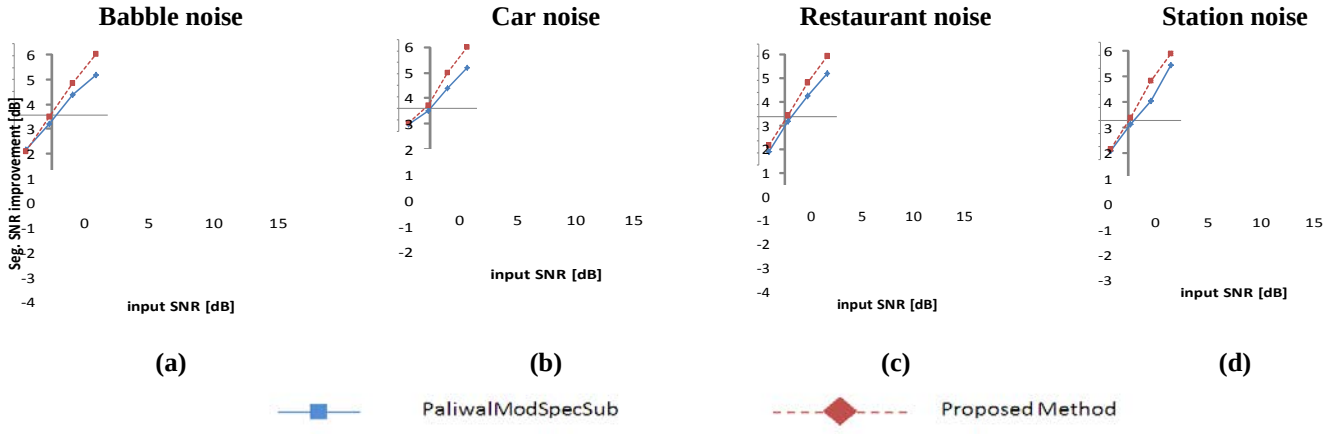

Fig. 3 : Mean Segmental SNR scores for a proposed approach compared to traditional Paliwals's ModSpecSub at different input SNR and noise type.

D. Composite objective measure:

The speech quality is also evaluated by composite objective measure (COM) [18]. Several composite objective quality measures are derived from multiple regression analysis. These measures include signal distortion (Csig), noise distortion (Cbak) and overall signal quality (Covl). Fig. 4 shows the averaged overall signal (Δ Covl) quality.
Overall signal quality is improved by 84.33% on average for airport noise at 0dB input SNR while an average improvement over 0-15 dB input SNR is about 28 % is reported.

E. Speech Intelligibility measure:

The improvement in the speech intelligibility of the proposed approach is evaluated with the help of STOI measure [19]. In addition to reducing time and costs compared to subjective listening experiments, STOI measure could also help to predict the intelligibility of the enhanced speech signal. In general, STOI shows high correlation with the intelligibility of noisy and enhanced speech signal resulting from noise reduction.
It is also evident from [19] that STOI shows the strong monotonic relation with the intelligibility scores of various listening tests. Fig. 5 shows improvement in average STOI scores of proposed OMSS as compared to the traditional ModSpecSub method.

F. Subjective evaluation:

The informal, subjective listening [20] quality test is conducted for assessing the quality of speech stimuli. Subjects: A group of 5 listeners (5 male, 4 female) with normal hearing and age group between 20 - 50 years participated in the listening test. The audio stimuli have been played using good quality head phone to this group, which are conducted in a sound proof room. Each listener is allowed to repeatedly play audio stimuli. Each listener is asked to rate the test audio stimuli as per the scale is shown in Table III. The average of subjective scores collected from score sheets of all participants is tabulated in Table IV.

Two NOIZEUS speech corpus sentences sp20 and sp25 of the different non-stationary background noise condition were applied to the subjective listening tests.

The first (sp20) sentence belongs to the female speaker and second (sp25) sentence belongs to the male speaker.

Table III MOS score

MOS score	Description	Level of distortion
5	Excellent	Imperceptible
4	Good	Perceptible, but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying, but not objectionable
1	Bad	Very annoying and objectionable

Table IV Results of subjective listening test in terms of MOS

	Spectral enhancement technique			
Noise Type	Proposed OMSS	Paliwal's ModSpecSub [11]	Berouti's [1]	Noisy stimuli
Airport	3.85	3.25	3.325	2.7
Babble	3.65	3.22	3.7	2.77
Car	4.062	3.85	3.675	2.9
Exhibition	4.05	3.75	3.575	2.95
Restaurant	3.95	3.47	3.4	2.85
Station	4.075	3.8	3.475	2.77

The MOS (mean opinion score) value of subjective listening in Table IV, show that the proposed approach gives better performance as compared to traditional spectral subtraction methods [1, 11].
In conventional single channel speech enhancement methods twinkling sounding noise called musical noise that can be quite annoying for the listener is observed. The speech synthesis in Paliwal's ModSpecSub method reported the annoying noise with speech temporal slurring whereas in proposed method the speech slurring in greatly reduced with little background noise.

G. Computational complexity:

The computational complexity of the proposed method with the traditional Modspecusub is found by running the

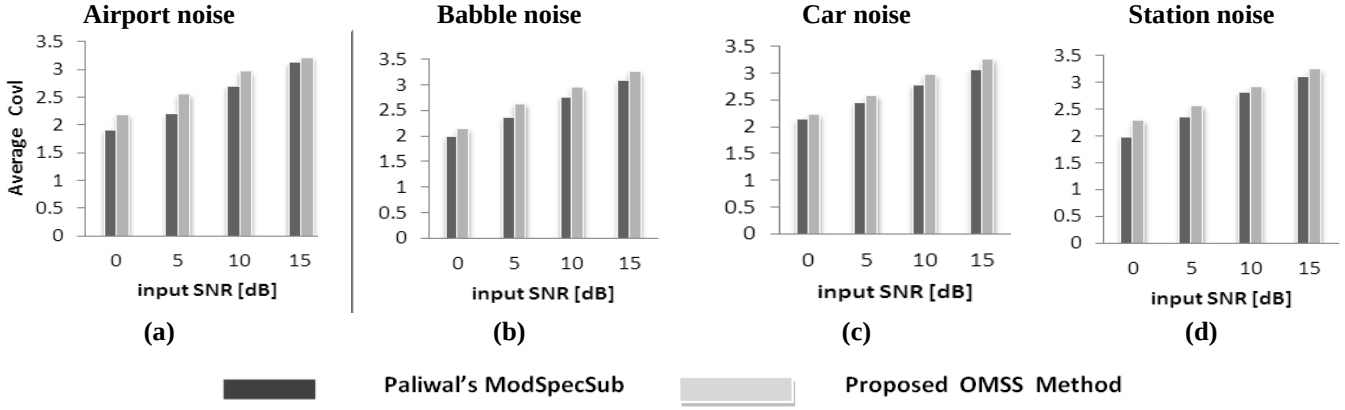

Fig. 4: *Average Overall signal quality (Covl) for different input noises and different input SNR.*

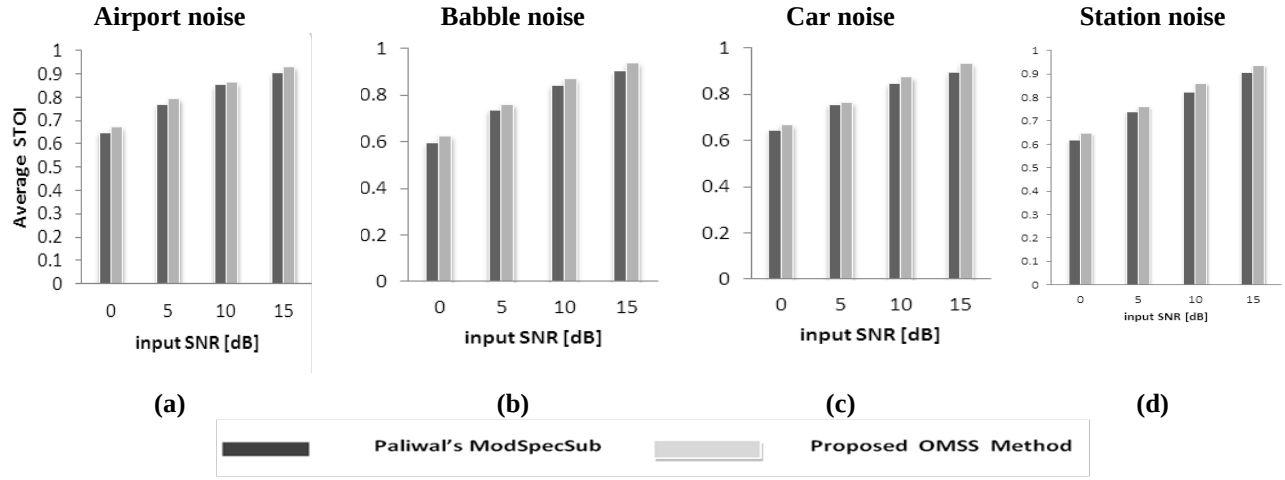

Fig. 5 *Average STOI measure for different non-stationary noise conditions at various input SNR.*

MATLAB simulations on a PC. The entire proposed approach is implemented on a computer system, build in Intel core i3 processor at the 2.4GHz clock frequency.

Table V Comparison of complexity

	ModSpecSub Method [11]		Proposed OMSS Method	
Normalized processing time	2.657		1	
	Calls	Total Time	Calls	Total Time
Hanning window	38916	7.856 s	2	0.04 s
Angle (Phase estimation)	38401	0.375 s	2	0.203 s
Specsub frame	38400	13.68 s	2048	0.170 s
Berouti [1]	38400	0.43 s	2048	0.05 s
specsub	512	14.94 s	--	--
repmat	2050	0.107 s	2	0.031 s
Noise estimate	--	--	1	3.336 s

We find the processing time required to run MATLAB simulation these methods. The computed values of processing time for ModSpecSub method are normalized with respect to processing time of OMSS method as shown in Table V.

One possible explanation would be that the ModSpecSub method utilizes VAD to update noise spectrum during speech absent, whereas OMSS method utilizes minimum statistics based spectral noise power estimation [13,14,15] from RMMS. The proposed method exhibits lower computational load compared to the ModSpecSub method. The comparison of complexity as shown in Table V is computed from profiler tool in Matlab. It gives the number of calls to an instruction along with its time. From Table V, normalized mean processing times for the proposed OMSS method is found to be improved.

H. Empirical waveform justification

Fig. 6 shows the speech stimuli of sp11 restaurant of NOIZEUS speech background noise at 5dB input SNR. The proposed OMSS approach synthesized time domain waveform shows the better closeness to the clean speech stimuli. It shows that the speech stimuli of proposed method follow the clean speech with very fewer distortions. It was also confirmed from the subjective listening test.

DISCUSSION

To compare the performance of the proposed approach in non-stationary environment to the existing modulation domain speech enhancement method, extensive experimental simulations are performed using a NOIZEUS speech corpus database. In the state of the art of speech enhancement methods proposed approach outperforms in terms of objective evaluation [17, 18] and subjective listening test for the different non-stationary environment. The proposed OMSS method achieves consistent improvement in speech quality across various input SNRs in terms LLR, WSS and subjective listening MOS scores as shown in Table I, 2 and 4 respectively.

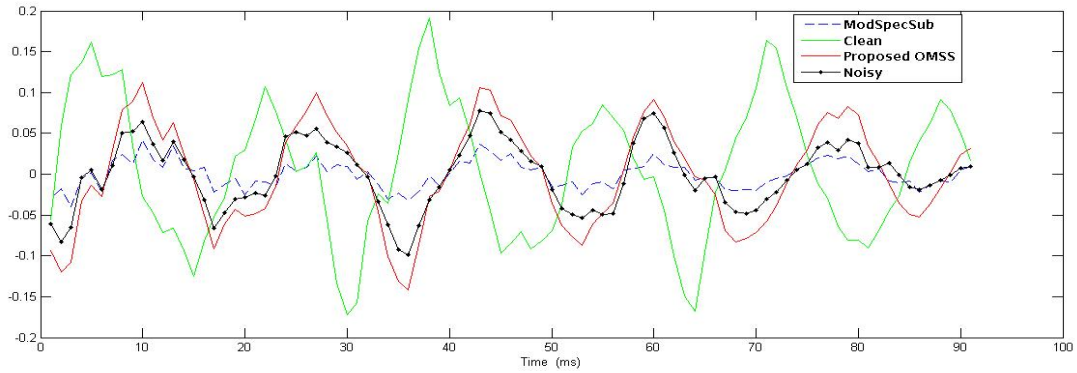

Fig. 6: *Speech temporal waveforms of utterance sp11 processed with the different speech enhancement methods along with clean utterance.*

The use of STOI [19] measure for evaluation of speech intelligibility has increased tremendously in the last decades. STOI objective intangibility measure reduces time and cost compare to the real listening test. STOI shows high correlation with the intelligibility of noisy signal and speech signal resulting from noise reduction. Improved speech intelligibility scoresare reported with the proposed OMSS method. It is observed informal listening test that Segmental SNR score is more robust over changing noise and different processing methods. The different acoustic and modulation frame durations were studied to enhance the noisy speech quality. The acoustic and modified modulation analysis frame duration 32 ms and 128 ms respectively, gives best objective scores as well as subjective scores for the proposed approach. We apply acoustic magnitude (alpha=1) and modulation magnitude in power form (i.e., alpha=2). From the informal listening test it is found that as we convert acoustic magnitude in square form (alpha=2) the background noise suppression is better but objective evaluation scores reduces.

CONCLUSION

We proposed a method for optimization of modulation domain signal processing using a traditional Analysis modification, synthesis. The proposed method is evaluated with different noise estimation techniques. The work presented in this paper explores AMS system along with the attributes of the modulation domain speech signal processing. The minimum statistics method of noise estimation method gives best objective and subjective scores among others. The performance of proposed approach has been evaluated by conducting extensive experiments using a speech corpus NOIZEUS database at different input SNR and various non-stationary noise conditions. We compare the traditional modulation spectral subtraction and modulation domain spectral subtraction with a proposed OMSS method with the several objective evaluation scores such as LLR, WSS, Segmental SNR and various composite objective measures. Also, the proposed approach achieves improved speech intelligibility assessed with STOI. Further, from the subjective listening experimental results, it is followed that the proposed approach outperforms than traditional modulation domain spectral subtraction in terms of perceived speech quality and intelligibility. Also, the computational load is reduced. It is improved by 57.13% as compared to traditional modulation spectral subtraction.

APPENDIX

Declarations

1	AMS	Analysis-modification-synthesis
2	AWGN	Additive white Gaussian noise
3	MMSE	Minimum Mean Square Error
4	OMSS	Optimized modulation spectral subtraction
5	MBSS	Multi Band Spectral Subtraction
6	MBCSS	Multiband complex spectral subtraction
7	ModSpecSub	Modulation spectral subtraction
8	SNR	Signal to noise ratio
9	WSS	Weighted Spectral Slope
10	LLR	Log Likelihood Ratio
11	SNRseg	Segmental SNR
12	STOI	Short-Time objective intelligibility
13	VAD	Voice activity detection

ACKNOWLEDGMENTS

The authors declare that there is no funding body involved in the parented work.

REFERENCES

[1] Berouti, M., Schwartz, R., Makhoul, J., Enhancement of speech corrupted by acoustic noise. In: Proc. IEEE Internat. Conf. Acoustics, Speech, and Signal Process. (ICASSP), 1979.Vol. 4. Washington, DC, USA, pp. 208–211.

[2]Boll S., 'Suppression of acoustic noise in speech using spectral subtraction'. IEEE Trans. Acoust. Speech Signal Process. 1979. ASSP-27 (2).

[3] Ephraim, Y., Malah, D.,: 'Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator'. IEEE Trans. Acoust. Speech Signal Process. 1984, ASSP-32 (6), pp. 1109–1121.

[4] Virag, N., 'Single channel speech enhancement based on masking properties of the human auditory system'. IEEE Trans. Speech Audio Process. 1999, 7 (2), pp. 126–137.

[5] Lim, J., Oppenheim, A. :'Enhancement and bandwidth compression of noisy speech'. Proc. IEEE 1979,67 (12), pp. 1586–1604.

[6] Kamath S., Loizou P.C.: 'A multi-band spectral subtraction method for enhancing speech corrupted by colored noise'. Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, Or-lando, Florida, USA, May 2002,vol. 4, pp. 4164–4164.

[7] SumanSamui, ChakrabartiI.,et.al, 'An improved single channel phase-aware speech enhancement technique for low SNR signal' IET Signal Processing, 2016,10(6),pp. 641 - 650.

[8] Griffin D., Lim J., 'Signal estimation from modified short-time Fourier transform'. IEEE Trans. Acoust. Speech Signal Process. 1984. ASSP-32 (2), pp. 236–243.

[9] Allen, J.,: 'Short term spectral analysis, synthesis, and modification by discrete Fourier transform'. IEEE Trans. Acoust. Speech Signal Process. 1977,25 (ASSP-3), pp. 235–238.

[10] R. E. Crochiere., 'A Weighted Overlap-Add Method of Short-Time Fourier Analysis/Synthesis'.IEEE Transaction on Acoustic,speech,and signal processing ,Vol. ASSP-28, NO. 1, Feb 1980, pp 99-102.

[11] KuldipPaliwal, Kamil Wo´jcicki, Belinda Schwerin, :'Single-channel speech enhancement using spectral subtraction in the short-time modulation domain', Speech Communication 2010, (52) pp. 450–475.

[12] 'NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms', http://ecs.utdallas.edu/loizou/speech/noizeus/, accessed 7 December 2015.

[13] Rainer Martin, 'Noise power spectral density estimation based on optimal smoothing and minimum statistics'. IEEE Trans. Speech and Audio Processing, 2001, 9 (5): pp. 504-512.

[14] Rainer Martin: 'Bias compensation methods for minimum statistics noise power spectral density estimation', Signal Processing, 2006, 86, pp. 1215-1229.

[15] Dirk Mauler and Rainer Martin, 'Noise power spectral density estimation on highly correlated data', Proceedings IWAENC, 2006

[16] Gerkmann T. &Hendriks R. C.,: 'Unbiased MMSE-Based Noise Power Estimation with Low Complexity and Low Tracking Delay', IEEE Trans Audio, Speech, Language Processing, 2012, 20, pp. 1383-1393.

[17] Loizou, P.,: Speech Enhancement: Theory and Practice. (Taylor and Francis, FL.2007)

[18] Hu, Y., Loizou P. C. 'Evaluation of objective quality measures for speech enhancement', IEEE Trans. Audio, Speech, Lang. Process., 2008, 16, (1), pp. 229-238.

[19] C.H.Taal, R.C.Hendriks, R.Heusdens, J.Jensen 'An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech', IEEE Transactions on Audio, Speech, and Language Processing, 2011vol. 19, no. 7, pp. 2125–2136.

[20] Hu, Y. and Loizou, P., :'Subjective evaluation and comparison of speech enhancement algorithms', Speech Communication, 2007,49, pp. 588-601.

[21] D. Klatt, "Prediction of perceived phonetic distance from critical band spectra," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 1982, vol. 7, pp. 1278 – 1281.

First A. Author

Mr. Pavan D. Paikrao has received B.E.(Electronics and Tele communication engineering) in 2009 and M. Tech (Electronics and Tele communication engineering) in 2011 from Dr. Babasaheb Ambedkar Technological University, lonere, Raigad, India. He is currently PhD student at Dr. Babasaheb Ambedkar Technological University, lonere, Raigad India. His research area includes ECG signal processing, speech signal processing.

Second B. Author

Dr. Sanjay L. Nalbalwar has received B.E. (Computer Science & Engineering) in 1990 and M.E. (Electronics) in 1995 from SGGS College of Engineering and Technology, Nanded, India. He has completed Ph.D. from IIT Delhi in 2008. He has around 20 years of teaching experience and is working as an Associate Professor & Head of Electronics & Telecommunication Engineering Department at Dr. Babasaheb Ambedkar Technological University, Lonere Raigad, Maharashtra State, India. His area of interest includes multirate signal processing and Wavelet, stochastic process modeling.