Speech Enhancement using a Modified Apriori SNR and Adaptive Spectral Gain Control

Ch.V.Rama Rao
Faculty Member, Dept.of ECE, Gudlavalleru Engineering College, Gudlavalleru-521356, AP, India.

M.B.Rama Murthy
Professor, Jayaprakash Narayan College of Engineering, Dharmapur, Mahabubnagar-509001, AP, India.

K.Srinivasa Rao
Principal, TRR College of Engineering, Pathancheru-502319, AP, India.

ABSTRACT
A new approach to single channel speech enhancement is proposed using a modified a priori SNR and spectral gain control. The proposed approach is first directed toward finding self adaptive averaging factor to estimate the apriori SNR. Next, spectral gain is reduced in order to suppress effects of the noise in the speech absent frames. Further, in the speech present frames, in order to reduce signal distortion, the spectral gain is controlled to be unity based on an SNR calculated by using a ridgeline spectrum. Finally, the original noisy speech is added to the estimated speech in a ratio is controlled by the long term averaged SNR of the estimated noise and the noisy speech. Computer simulations by using speech signals, the white noise, the car noise and the babble noise have been carried out using several available methods and the proposed method. It is observed that there is improvement in speech quality by the proposed method.

Key words: MMSE estimator, apriori SNR, spectral gain.

1. INTRODUCTION
Mobile communication systems and various speech processing systems have been found their way in our everyday life through their vivid use in voice communication, speech and speaker recognition, aid for the hearing impaired, and numerous other applications [1]. However, in most of the cases the ambient environment is noisy which degrades the performance of the speech processing systems drastically. Therefore, speech enhancement has been a challenging topic of research for many years.

A specyral suppression technique is a hopeful approach, and many kinds of noise spectral suppressors have been developed for mobile phones. Several methods for estimating a spectral gain, which is used to suppress the noise spectrum, have been proposed. They include MMSE-STSA [2], MMSE- LSA [3] and Joint MAP [4]. Performance of the noise suppressor based on the spectral suppression technique is highly dependent on apriori SNR to reduce musical noise. However, the enhanced speech is accompanied by unpleasant musical noise artefact which is characterized by tones with random frequencies. Apart from being extremely annoying to the listeners, the musical noise also hampers the performance of the speech coding algorithms to a great extent.

This paper is directed toward finding an improved estimate of the apriori SNR, which in turn has the affect of reducing the musical noise produced by the conventional methods. It has been shown in [5] that the key point behind the reduction of musical noise by the MMSE estimator [2] is the use of apriori SNR. In this work, a method for calculating the time-frequency varying averaging factor, used in estimating the apriori SNR, is proposed. Furthermore, the spectral gain is adaptively controlled depending on whether the speech absent or present frames. The speech absent and present frames are precisely detected by using a voice activity detector [5].

2. SPECTRAL SUPPRESSION TECHNIQUE
Let $s(n)$, $d(n)$ and $y(n)$ be a clean speech, noise and their mixed signal, that is noisy speech, respectively.

$$y(n) = s(n) + d(n) \tag{1}$$

Let $S(m,k)$, $D(m,k)$ and $Y(m,k)$ be the FFT of $s(n)$, $d(n)$ and $y(n)$, respectively. m is a frame number, and k is a frequency number. They are related by

$$Y(m,k) = S(m,k) + D(m,k) \tag{2}$$

The spectral gain $G(m,k)$ at the m^th^ frame is calculated, and the noise spectrum is suppressed by

$$\hat{S}(m,k) = G(m,k)Y(m,k) \tag{3}$$

$G(m,k)$ is calculated by MMSE-STSA [2] and MMSE-LSA [3] using estimated prior and posterior SNRs. A prior SNR $\xi(m,k)$, which is a ratio of the clean speech spectrum and the noise spectrum, and a posterior SNR $\gamma(m,k)$, which is a ratio of the noisy speech spectrum and the noise spectrum, are expressed by

$$\xi(m,k) = \frac{\lambda_s(m,k)}{\lambda_d(m,k)} \tag{4}$$

$$\gamma(m,k) = \frac{|Y(m,k)|^2}{\lambda_d(m,k)} \tag{5}$$

$$\lambda_s(m,k) = E\{|S(m,k)|^2\} \tag{6}$$

$$\lambda_d(m,k) = E\{|D(m,k)|^2\} \tag{7}$$

In the above equations, only the noisy speech spectrum $|Y(m,k)|^2$ is available, and the other spectra should be estimated. The prior SNR $\xi(m,k)$ can be estimated by [2].

$$\hat{\xi}(m,k) = \alpha\frac{|\hat{S}(m-1,k)|^2}{\lambda_d(m-1,k)} + (1-\alpha)P[\gamma(m,k)-1] \tag{8}$$

Where $\hat{S}(m-1,k)$ denotes the amplitude estimate of the kth speech spectral component, $\lambda_d(m-1,k)$ is the estimate of variance of the kth noise spectral component in the $(m-1)^{th}$ analysis frame, the operator $P[\cdot]$ denotes half wave rectification, and α denotes an averaging parameter. Considering the maximum-likelihood estimate of the apriori SNR, we have $\xi(m,k) = E\{\gamma(m,k-1)\}$.

3. PROPOSED MODIFICATION OF APRIORI SNR

In the expression of $\hat{\xi}(m,k)$ given by (8), the choice of α is critical. In general, α is given a value very close to 1. It has been shown [4] that the closer the value of α is to 1, the lesser is the musical noise, but there is more "transient distortion" to the resulting signal. Balancing these two effects, reported results in the literature usually set α a constant value in the range 0.95-0.99 with a few exceptions [6][7]. But using a constant α has certain drawbacks. Consider an example as a test case where α =0.98, and the posterior SNR shows a pulse like behaviour, i.e., for $m < m_1$ and $m > m_2$, it is very low as compared to its values in the interval m_1 to m_2, where m_1 and m_2 denote, respectively the frames of rising and falling edges. At m_1, a signal component suddenly goes high such that $\gamma(m_1,k) \geq \gamma(m_1-1,k)$. Since $\hat{\xi}(m,k)$ contains 98% of the previous frame estimated SNR, it will fail to respond to this change. Rather, $\hat{\xi}(m,k)$ will rise slowly and ultimately begin to follow posterior SNR in this high SNR region (m_1 to m_2) with some delay. Similarly at m_2, $\hat{\xi}(m,k)$ fails to respond to the abrupt downfall of posterior SNR and only after a certain delay converges to the low SNR level. Therefore, it will be logical to use a much smaller value of α in these transitional areas. In [7], an adaptation scheme for α has been defined based on the assumption that the additive noise is stationary and the noise energy does not change significantly from frame to frame. A formulation using the frame energy is given by

$$\alpha_m = \sqrt{\left(1 - \frac{|E_m - E_{m-1}|}{\max(E_m, E_{m-1})}\right)} \tag{9}$$

Where $E_m = \sum_k |Y(m,k)|^2$. Clearly, the above expression for α is based on intuitive arguments. In this work, we develop an expression for self-adaptive α based on MMSE criterion to account for the abrupt changes in the speech spectral amplitude. The proposed modification in the estimation of apriori SNR is given by

$$\hat{\xi}(m,k) = \alpha(m,k)\tilde{\xi}(m-1,k) + (1-\alpha(m,k))P[\gamma(m,k)-1] \tag{10}$$

where $\tilde{\xi}(m-1,k) = \left|\hat{S}(m-1,k)\right|^2 / \lambda_d(m-1,k)$.

Since the estimate $\hat{\xi}(m,k)$ given by (10) should actually be as close as possible to apriori SNR $\xi(m,k)$, we propose an MMSE estimator for $\alpha(m,k)$ that minimizes the error

$$J_\alpha = E\left\{\left(\hat{\xi}(m,k) - \xi(m,k)\right)^2 / \tilde{\xi}(m-1,k)\right\} \tag{11}$$

Given $\tilde{\xi}(m-1,k)$. Substituting (10) into (11), an expression for J_α can be obtained as

$$J_\alpha = \alpha^2(m,k)\left(\tilde{\xi}(m-1,k) - \xi(m,k)\right)^2 + (1-\alpha(m,k))^2(\xi(m,k)+1)^2. \tag{12}$$

Note that in the above derivation $E\left\{(\gamma(m,k)-1)^2\right\} = 2\xi^2(m,k) + 2\xi(m,k) + 1$ is substituted. This result is obtained assuming that both noise and speech spectral coefficients are statistically independent zero mean complex Gaussian random variables, and using the relation $E\left\{|S(m,k)|^4\right\} / \lambda_d^4(m,k) = 2\xi^2(m,k)$, which follows from the definition of the fourth moment with the assumption that speech spectral amplitude $|S(m,k)|$ has a Rayleigh distribution [8].

Now equation $\partial J_\alpha / \partial\alpha(m,k)$ to zero, we obtain an expression for optimum $\alpha(m,k)$ as

$$\alpha^{opt}(m,k) = \frac{1}{1 + \left(\dfrac{\xi(m,k) - \tilde{\xi}(m-1,k)}{\xi(m,k+1)}\right)^2}. \tag{13}$$

As $\xi(m,k)$ is unknown, (13) cannot be used directly. Nevertheless, an approximate value of $\alpha(m,k)$ can be obtained substituting $\bar{\xi}(m,k) = P[\gamma(m,k)-1]$ for $\xi(m,k)$ in (13). This is a reasonable substitution as $E\left\{\bar{\xi}(m,k)\right\} \cong \xi(m,k)$. If posterior SNR over a region shows uniform variation, $\alpha(m,k)$ will attain a value close to 1. For any abrupt change, $\alpha(m,k)$ attains a lower value enabling $\hat{\xi}(m,k)$ to respond to that change more suitably.

4. ADAPTIVE SPECTRAL GAIN

In the spectral suppression technique, there exists a trade-off the large residual noise due to a lack of noise suppression and speech distortion caused by over suppression. In this work, an adaptive control method for spectral gain is proposed. In this work, the noise spectrum is more precisely estimated in the speech absent frame. For this purpose, a voice activity detector (VAD) is applied. Essentially, in order to estimate the noise spectrum in the nonstationary environment, the VAD using the spectral entropy [5] is employed.

4.1 Modification of spectral gain

In the speech absent frames, the spectral gain is more reduced in order to well suppress the noise spectrum.

$$G(m,k) = \begin{cases} G_{\sup} G(m,k) & \text{in speech absent frames} \\ G(m,k) & \text{in speech present frames} \end{cases} \tag{14}$$

Next, in order to reduce speech distortion due to over suppression, the minimum values G_{floor} and $G_{\min}$ in the speech present frames are set to be larger than those in the speech absent frames.

4.2 Adaptive control of spectral gain based on speech spectrum estimation

A ridgeline spectrum and an offset-SNR are estimated as follows [5]:

a) The peak amplitude $|Y_{\max}(k)|$ is updated by using the noisy speech spectrum $|Y(m,k)|$ as follows:

If $|Y_{\max}(k)| < |Y(m,k)|$, then $|Y_{\max}(k)| = |Y(m,k)|$

b) When the noisy speech spectrum $|Y(m,k)|$ is close to the peak amplitude $|Y_{\max}(k)|$, the ridgeline spectrum $|Y_{edge}(m,k)|$ is updated as follows: If $|Y(m,k)| > \beta_{\max}|Y_{\max}(k)|$, then $|Y_{edge}(m,k)| = \mu_r|Y_{edge}(m-1,k)| + (1-\mu_r)|Y(m,k)|$

c) An offset SNR (SNR_{offset}) is calculated based on a ratio of the ridgeline spectrum and the estimated noise spectrum, which are averaged in the frequency domain.

$$SNR_{offset} = \frac{\sum_{k=1}^{2M}(1-\alpha_{offset})|Y_{edge}(m,k)|}{\sum_{k=1}^{2m}\sqrt{\lambda_d(m,k)}}$$

$\beta_{\max}$, μ_r and α_{offset} are positive constants less than 1. SNR_{offset} is the posterior SNR by which the speech distortion does not occur. In the proposed method, SNR_{offset} is used as follows: in the speech present frames, which are detected by the VAD, if the posterior SNR is higher than SNR_{offset}, then the spectral gain is set to be $G(m,k)$=1.

4.3 Adding Original Noisy Speech

In the spectral suppression technique, the noise spectrum can be suppressed, however, at the same time, the speech itself is usually distorted. By adding the original noisy speech to the estimated speech, the speech becomes more natural at the expense of the remaining noise.

In this work, a rate of adding the original noisy speech is controlled by the long term average $\gamma_{avg}(m,k)$ of the posterior SNR as follows:

$$\gamma_{avg}(m,k) = \alpha_{avg}\gamma(m,k) + (1-\alpha_{avg})\gamma(m-1,k) \tag{15}$$

$$\hat{G}(m,k) = \begin{cases} n_1 + (1-n_1)G(m,k); & SNR_{th1} < \gamma_{avg}(m,k) \\ n_2 + (1-n_2)G(m,k); & SNR_{th2} < \gamma_{avg}(m,k) < SNR_{th1} \\ G(m,k); & otherwise \end{cases} \tag{16}$$

$$\hat{s}(n) = IFFT[\hat{G}(m,k)|Y(m,k)|e^{j\theta(m,k)}] \tag{17}$$

$\hat{s}(n)$ is the estimated speech in the mth frame. The parameters are determined by experience as follows: n_1=0.75, n_2=0.15, SNR_{th1}=12 dB and SNR_{th1}=5dB.

5. RESULTS AND DISCUSSIONS

In the experiments, noisy speech signals are obtained by adding a clean speech signal with white, babble (speech-like), airport, train, street and car noise signals which are extracted from the NOIZEUS database [9]. Four SNR levels, including 0 dB, 5dB, 10dB and 15dB, are used to evaluate the performance of a speech enhancement system. The proposed, MMSE STSA estimator (method1) and the power spectral subtraction (method 2) is also conducted for comparisons.

The performance of the proposed method is evaluated by using the following objective measures

5.1 Time-domain SNR measures

The time-domain segmental SNR (SNR $_{seg}$) measure [10] was computed is given by

$$SNR_{seg} = \frac{10}{M}\sum_{m=0}^{\omega-1}\log\frac{\sum_{n=N_m}^{N_m+N-1}s^2(n)}{\sum_{n=N_m}^{N_m+N-1}\left(s(n)-\hat{s}(n)\right)^2} \tag{18}$$

Where $s(n)$ the input is (clean) signal, $\hat{s}(n)$ is the processed (enhanced) signal, N is the frame length and M is the number of frames in the signal. Table 1 presents the performance comparisons in terms of the average segmental (Avg. SegSNR) improvement which is computed by subtracting the Avg.Seg SNR of noisy speech from that of enhanced speech. All of the speech enhancement algorithms provide more Avg.SegSNR improvement in low-SNR inputs. The best performance is obtained in the case of white noise corruption. It is due to the fact that the spectra of white noise widely spread over frequency subbands.

In the cases of low-SNR inputs (0dB and 5dB), the proposed approach significantly outperforms the MMSE STSA estimator (method1) and the power spectral subtraction (method 2). It is attributed to the fact that the proposed method is better able to reduce the amounts of residual noise than the other two methods.

5.2 Log-likelihood ratio (LLR) measure

The performance of the proposed method was evaluated The LLR measure [11] for each 20-ms speech frame was computed as follows:

$$d_{LLR} = \log_{10}\left(\frac{a_y R_s a_y^T}{a_s R_s a_s^T}\right) \tag{19}$$

Where a_s and R_s are the linear prediction coefficient vector and autocorrelation matrix of the original(clean) speech frame respectively, and a_y is the linear prediction coefficient vector of the enhanced speech frame. The LLR is a spectral distance measure which mainly models the mismatch between the formats of the original and enhanced signals. The mean LLR value was obtained by averaging the individual frame LLR values across the sentence. The LLR results are tabulated in Table 2, smaller spectral distance values (LLR) were obtained by the proposed method.

Figure 1 demonstrates an example of waveform plots for comparisons. A speech signal uttered by a male speaker was corrupted by babble noise with 5 dB SNR. Since babble noise is a speech like noise and how to remove this kind of noise from a corrupted noisy speech signal is difficult. Comparing the waveforms of enhanced speech shown in Figure's 1(c), 1(d) and 1(e), the proposed method significantly improves the performance of the speech enhancement system in removing background noise. Although the proposed method significantly reduces the amounts of residual noise in speech-pause regions, the enhanced

Table 1. Comparisons of Avg. SegSNR improvement

Noise type	SNR (dB)	Enhancement method		
		Method1	Method2	Proposed Method
White	0	0.02	0.01	5.97
	5	0.23	0.14	7.07
	10	0.79	0.57	8.04
	15	1.62	1.32	8.77
Babble	0	0.9	1.09	4.43
	5	1.43	1.02	3.24
	10	0.04	-0.15	2.34
	15	-1.2	-3.15	-0.88
Airport	0	0.14	1.65	4.26
	5	0.5	0.84	4.16
	10	-2.39	-3.01	-0.8
	15	-1.51	-2.62	-0.25
Car	0	1.54	0.99	5.21
	5	1.66	1.41	5.07
	10	1.04	0.28	2.91
	15	-1.01	-2.55	-0.25
Train	0	1.44	1.45	3.83
	5	2.13	1.24	4.05
	10	-0.42	-0.84	1.79
	15	-1.61	-3.74	-1.68
Street	0	2.26	1.98	4.02
	5	1.25	0.79	3.20
	10	1.38	-0.13	2.74
	15	-1.66	-3.32	-1.00

Table 2. LLR values

Noise type	SNR (dB)	Enhancement method		
		Method1	Method2	Proposed method
White	0	2.66	2.59	1.78
	5	2.61	2.68	1.90
	10	2.60	2.56	1.68
	15	2.47	2.36	1.41
Babble	0	1.22	1.16	0.87
	5	0.97	0.91	0.53
	10	0.70	0.68	0.33
	15	0.41	0.60	0.32
Airport	0	1.23	1.10	1.02
	5	0.77	0.84	0.47
	10	0.61	0.71	0.46
	15	0.44	0.62	0.26
Car	0	1.22	1.15	1.02
	5	0.89	0.83	0.45
	10	0.58	0.66	0.37
	15	0.46	0.59	0.28
Train	0	1.37	1.27	1.07
	5	0.98	0.95	0.54
	10	0.85	0.83	0.55
	15	0.67	0.83	0.42
Street	0	1.32	1.20	1.25
	5	1.15	1.07	0.57
	10	0.68	0.68	0.40
	15	0.54	0.66	0.28

speech signal does not been severely deteriorated during speech-dominant regions. Therefore, the speech quality can be maintained at an acceptable level.

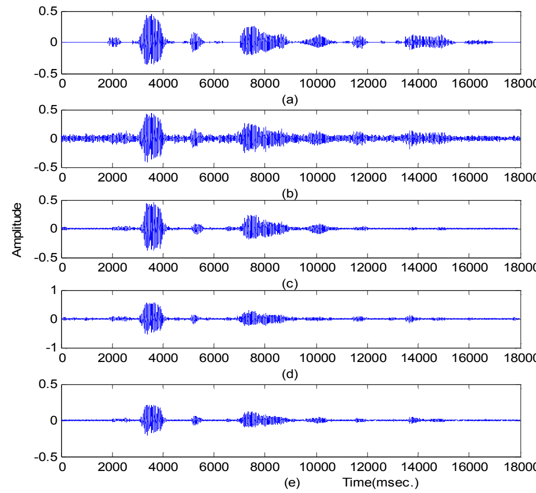

Fig 1: Time plots of (a) clean speech (b) noisy speech corrupted by babble noise with 5 dB SNR, (c) enhanced speech using MMSE-STSA estimator, (d) enhanced speech using power spectral subtraction and (e) enhanced speech using proposed method

Figure 2 shows the spectrograms of a speech signal uttered by a male speaker corrupted by babble noise with 5 dB SNR. Observing the spectrograms of enhanced speech during speech-pause regions, the proposed method shown in Figure 2(e) is better able to remove background/residual noise than the other two methods shown in Figure's 2(c) and 2(d). An informal subjective listening test also reveals that the enhanced speech produced by the proposed method sounds like less annoying that that produced by the MMSE spectral amplitude estimator and the power spectral subtraction methods.

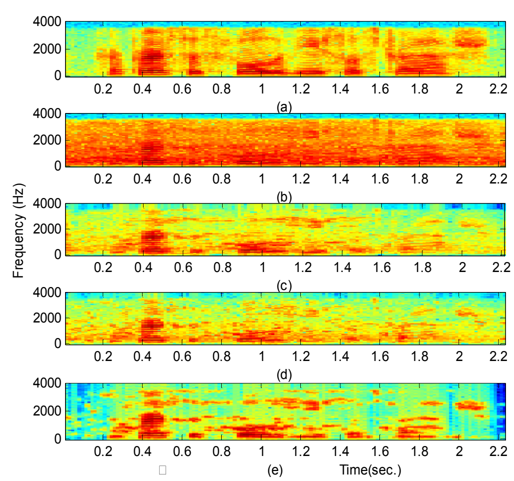

Fig 2: Spectrograms of (a) clean speech, (b) noisy speech corrupted by babble noise with 5 dB SNR, (c) enhanced speech using MMSE-STSA estimator, (d) enhanced speech using power spectral; subtraction and (e) enhance speech using proposed method.

6. CONCLUSION

In this work, an effort has been made to develop an optimal expression for the time-frequency verying averaging factor in the MMSE sense to estimate more accurately the apriori SNR and the adaptive spectral gain control for noise suppression. The average segmental SNR and the log-likelihood ratio (LLR) are evaluated by using the speech and several noises. The proposed metgod is superior to the conventional methods in all noise environments.

7. REFERENCES

[1] J.S.Lim and A.V.Oppenheim, "Enhancement and bandwidth compression of noisy speech," Proc. IEEE, vol.67, pp. 1586-1604, Dec. 1979.

[2] Y.Ephraim and D.Malah, "Speech enhancement using minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Vol. ASSP-32, no. 6, pp. 1109-1121, Dec. 1984.

[3] Y.Ephraim, and D.Mallah, "Speech enhancement using minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Vol.ASSP-33, no. 2, pp. 443-445, April 1985.

[4] O.Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," IEEE Trans. Speech Audio Process., vol.2, no.2, pp. 345-349, April 1994.

[5] B.F.Wu and K.C.Wang, "Noise spectrum estimation with entropy-based VAD in non-stationary environments," IEICE (Japan) Trans. Fundamentals, vol. E89-A, no. 2, Feb. 2006.

[6] C.Beaugeant and P.Scalart, "Noise reduction using perceptual spectral change," in Proc. EUROSPEECH Conf., vol. 6, pp. 2543-2546, 1999.

[7] I.Y.Soon and S.N.Koh, "Low distortion speech enhancement," Proc. Inst. Elect. Eng., Vis. Image Signal Process., vol. 147, pp. 247-253, 2000.

[8] B.L.Sim, Y.C.Tong, J.S.Chang and C.T.Tan, "A parametric formulation of the generalized spectral subtraction method," IEEE Trans. Speech Audio Processing, vol. 6, pp. 328-337, July 1998.

[9] Yi Hu and Philipos C.Loizou, "Evaluation of objective quality measures for speech enhancement", IEEE Trans. On Audio, Speech and Language Processing, vol.16, no.1, pp. 229-238, Jan.2008.

[10] Ma, Jianfen, Yi Hu and Philipos C.Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-important function", J. Acoust. Soc. Am., Vol. 125, No. 5, pp. 3387-3405, 2009.

[11] Quackenbush.S, Barnwell. T and Clements. M, "Objective measures of speech quality," Prentice Hall, Englewood Cliffs, NJ, 1988.

AUTHORS PROFILE

1. CH.V.Rama Rao was born on 15th June 1979. He obtained his B.E., in ECE from Nagarjuna University in 2000, M.Tech. in Digital System and Computer Electronics from JNTU, Hyderabad in 2007 and he is pursuing Ph.D at a JNTU, Hyderabad. He has a teaching experience of over 10 years and research experience of over 2.5 years. He published ten papers in international and national journals and conferences. His areas of interest are Speech Processing, Signal Processing and Communications.

2. M B Ramamurthy is in the academic filed for the past 36 years. Currently he is Professor ECE and Dean Academics in Jayaprakash Narayan College of Engineering, Mahabubnagar India 509001. He has 66 publicaions to his credit in International Journals and Conferences. He is senior member IEEE, life fellow IETE, life member IEI and life member ISOI. His areas of interest are Speech Processing, Signal Processing and Communications.

3. K.Srinivasa Rao was born on 16th January 1962. He obtained his B.E., in ECE from Andhra University in 1985, M.Tech. in Digital Systems Engineering from Osmania University in 1993 and Ph.D., in ECE from Andhra University in 2000. He has a teaching experience of over 25 years and research experience of over 8 years. He published 31 papers in various international and national journals and conferences. He is member of IEEE, Fellow IETE and life member of ISTE, SEMCE (I) and MBMESI. His areas of interest are Digital Electronics, Microwave Communications, Communication Systems, Antennas & Propagation, VLSI, Embedded Systems and Image Processing.