World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:3, No:3, 2009

Speech Coding & Recognition

M. Satya Sai Ram, P. Siddaiah, and M. Madhavi Latha

***Abstract*—**This paper investigates the performance of a speech recognizer in an interactive voice response system for various coded speech signals, coded by using a vector quantization technique namely Multi Switched Split Vector Quantization Technique. The process of recognizing the coded output can be used in Voice banking application. The recognition technique used for the recognition of the coded speech signals is the Hidden Markov Model technique. The spectral distortion performance, computational complexity, and memory requirements of Multi Switched Split Vector Quantization Technique and the performance of the speech recognizer at various bit rates have been computed. From results it is found that the speech recognizer is showing better performance at 24 bits/frame and it is found that the percentage of recognition is being varied from 100% to 93.33% for various bit rates.

***Keywords*—**Linear predictive coding, Speech Recognition, Voice banking, Multi Switched Split Vector Quantization, Hidden Markov Model, Linear Predictive Coefficients.

International Science Index, Electronics and Communication Engineering Vol:3, No:3, 2009 waset.org/Publication/3722

I. INTRODUCTION

THIS paper takes the advantage of voice banking application and examined the performance of a speech recognizer in an Interactive voice response system for the coded output obtained by using Multi switched split vector quantization technique (MSSVQ) at various bit rates. MSSVQ has already been proved that it has better Spectral distortion performance, less Computational complexity and less Memory requirements when compared to other product code vector quantization techniques. So this paper uses MSSVQ as the vector quantization technique for coding.

Voice Banking is a tremendous telephone banking service that makes the user to be in touch with his account information and other banking services 24 hours a day 365 days a year by making a simple phone call. In voice banking customers can speak their choices, or can use a touch tone keypad to enter selections.

The speech techniques involved in voice banking are the speech coding, speech enhancement and speech recognition. This paper investigates the performance of a speech recognizer using hidden markov model (HMM) technique ([1],[2],[3]) for the coded outputs obtained by using a hybrid vector quantization technique. The hybrid vector quantization technique used for coding is the Multi Switched Split vector quantization (MSSVQ) technique ([4],[5],[6],[7]). The speech parameters used for coding are the line spectral frequencies (LSF) ([8],[9],[10]) so as to ensure the filter stability, the codebooks used for coding are generated by using the Linde Buzo Gray (LBG) algorithm [11] the generation of the codebooks is a tedious and time consuming process requiring large amounts of memory for generation and storing purposes, the memory required for the generation of the codebooks increases with the number of training vectors number of samples per vector and bits used for codebook generation.

The speech recognition technique used for recognition is the hidden markov model technique. HMM is a collection of various statistical modeling techniques, in which the transition probability matrix is estimated by using the Baum Welch algorithm ([1],[2]), the emission matrix is generated by using the K-means clustering algorithm and is estimated by using the Baum Welch algorithm. The Viterbi algorithm can also be used for the estimation of the transition and emission matrices. For a given sequence the most likely sequence path is estimated by using the Viterbi algorithm ([1],[2]), from which probability of a particular sequence is estimated by using the forward algorithm or the backward algorithm.

The aim of this article is to investigate the performance of the speech recognizer using HMM for a coded output obtained by using multi switched split vector quantization technique at different bit rates. The speech parameters that can be used for recognition are the Linear predictive coefficients (LPC) and Mel Cepstrum coefficients (MFCC) .In this paper LPC coefficients were used for recognition and Line spectral frequencies were used for coding To improve the performance of recognition energy, delta and acceleration coefficients must be used but in this paper they were not used because if they were used the generation of codebooks during coding becomes a problem.

II. SPEECH CODING AND RECOGNITION

This paper is intended for voice banking application, so it requires the technology of speech coding and recognition. The enhancement technique used is the Spectral subtraction technique ([11],[12],[13]). The coding technique used is the Multi Switched Split Vector Quantization technique (MSSVQ). The recognition technique used is the Hidden Markov model technique. The steps involved in speech coding and recognition intended for voice banking are

- Firstly the silence part of the speech signal is removed by using the voice activation and detection technique and next the channel noise included in the speech signal must be removed by using an enhancement technique.

M. Satya Sai Ram, Department of ECE, R.V.R & J.C. College of Engineering,Guntur-522019,A.P, India (e-mail: m_satyasairam@yahoo.co.in).
P. Siddaiah, Department of ECE, K.L College of Engineering, Guntur-522502, A.P, India (e-mail: siddaiah_p@yahoo.com).
M. Madhavi Latha Department of ECE J.N.T.U College of Engineering Hyderabad-500072, A.P, India (e-mail: mlmakkena@yahoo.com).

- Secondly the speech signal must be coded by using the MSSVQ technique.
- Thirdly the coded output with added channel noise must be enhanced by using the spectral subtraction technique.
- Next the enhanced speech signal must be given to a voice bank recognizer so as to recognize the coded output.
- Finally the percentage of recognition was computed as a measure of the recognition accuracy.

By using these speech techniques it is found that the recognition accuracy is being varied from 100% to 93.33% for the coded outputs at different bit rates.

III. MULTI SWITCHED SPLIT VECTOR QUANTIZATION

In MSSVQ for a particular switch the generation of codebooks at different stages is shown in Fig. 1.

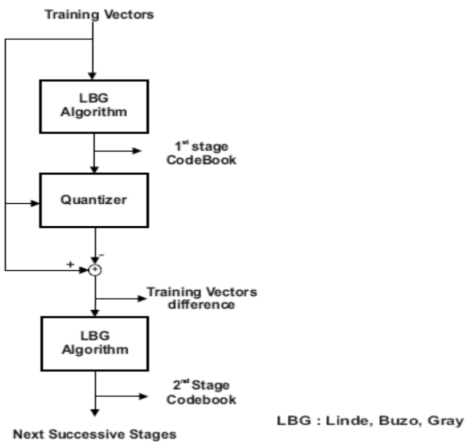

Fig. 1 Codebook Generation at different stages

- Initially the codebook at the first stage is generated by using the Linde, Buzo and Gray (LBG) [14] algorithm with the training vectors set as an input.
- Secondly the training difference vectors are extracted from the input training vectors set and the quantized training vectors of the first stage.
- Finally the training difference vectors are used to generate the codebook of the second stage.

This procedure is continued for the required number of stages and the number of codebooks to be generated will be equal to the number of stages used for quantization.

A p x m x s MSSVQ is shown in Fig. 2, where p corresponds to the number of stages, m corresponds to the number of switches, and s corresponds to the number of splits.

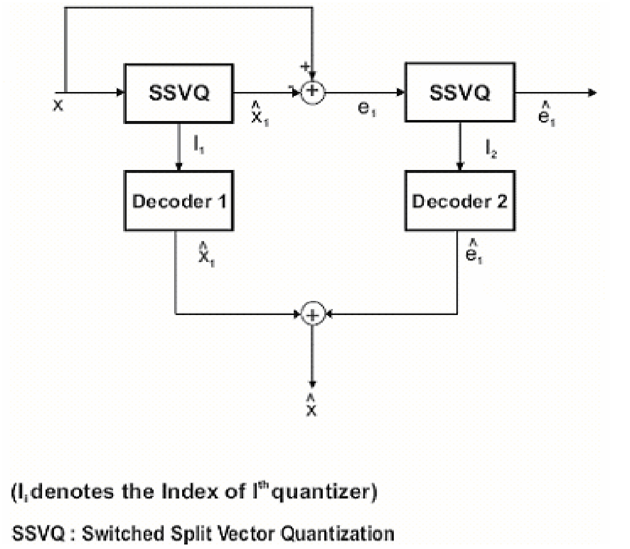

Fig. 2 Block Diagram of MSSVQ

- Each input vector x that is to be quantized is applied to SSVQ at the first stage so as to obtain the approximate vectors at each codebook of the first stage.
- Extract the approximate vector with minimum distortion from the set of approximate vectors at the first stage i,e. $\hat{x}_1 = Q[x_1]$.
- Compute the error vector resulting at the first stage of quantization and let the error vector be, $e_1 = x_1 - \hat{x}_1$.
- The error vector at the first stage is given as an input to the second stage so as to obtain the quantized version of the error vector $\hat{e}_1 = Q[e_1]$.

This process is continued for the required number of stages. Finally the decoder takes the indices, I_i, from each stage and adds the quantized vectors at each stage so as to obtain the reconstructed vector $\hat{x}$ given by $\hat{x} = Q[x_1] + Q[e_1] + Q[e_2] + \ldots\ldots$. Where $Q[x_1]$ is the quantized input vector at the first stage, $Q[e_1]$ is the quantized error vector at the second stage and $Q[e_2]$ is the quantized error vector at the third stage and so on.. As this process involves the quantization of the error vectors and summing of the error vectors with the approximate vector at the first stage the spectral distortion performance can be greatly improved when compared to SSVQ and SVQ.

IV. SPECTRAL DISTORTION

In order to objectively measure the distortion between a coded and uncoded LPC parameter vector, the spectral distortion is often used in narrow band speech coding. For the i^{th} frame the spectral distortion (in dB), SD_i, [5] is defined as

$$SD_i = \sqrt{\frac{1}{(f_2 - f_1)} \int_{f_1}^{f_2} \left[10\log_{10} x_i(f) - 10\log_{10} \hat{x}_i(f)\right]^2 df} \; (dB) \qquad (1)$$

Where F_S is the sampling frequency and $x_i(f)$ and $\hat{x}_i(f)$ are the LPC power spectra of the uncoded and coded i^{th} frame, respectively. f is the frequency in Hz, and the frequency range is given by f_1 and f_2. the frequency range used in practice is 0-4000Hz. The average spectral distortion SD is given by

$$SD = \frac{1}{N}\sum_{n=1}^{N} SD_i \qquad (2)$$

The conditions for transparent speech from narrowband LPC parameter quantization are.

- The average spectral distortion (SD) must be less than or equal to 1dB.
- There must be no outlier frames having a spectral distortion grater than 4dB.
- The no of outlier frames between 2 to 4dB must be less than 2%.

V. Results

Tables I to IV gives the probability of recognizing an utterance ONE at bit rates 24, 23, 22, 21. From tables it is observed that the recognition accuracy is being varied from 100% to 93.33% for different bit rates and it is found that the recognition accuracy is good at 24 and 23 bits/frame. The reason for choosing multi switched split vector quantization technique is that it is having better spectral distortion performance, less computational complexity and less memory requirements when compared to other product code vector quantization techniques which can be observed from Tables V to VIII. As a result the cost of the product will be less when using MSSVQ and can have better marketability. The decrease in spectral distortion, complexity and memory requirements for MSSVQ can also be observed from Fig's 3 to 5. The spectral distortion is measured in units of decibles (dB), computational complexity is measured in units of kflops/frame, and memory requirements are measured in units of floats.

TABLE I
PROBABILITY OF RECOGNIZING A WORD ONE AT 24 BITS/FRAME BY USING MSSVQ

NAME	PROBABILITY
ZERO	-20.5326
ONE	-16.9179
TWO	-18.6235
THREE	-18.6513
FOUR	-19.4956
FIVE	-21.7565
SIX	-17.0356
SEVEN	-19.3630
EIGHT	-19.4613
NINE	-19.6206
TEN	-18.7590
YES	-17.7631
NO	-19.1707
SUCESSFULL	-20.1300
UNSUCESSFULL	-22.7260
% RECOGNITION	100%

TABLE II
PROBABILITY OF RECOGNIZING A WORD ONE AT 23 BITS/FRAME BY USING MSSVQ

NAME	PROBABILITY
ZERO	-21.1627
ONE	-14.0445
TWO	-17.7641
THREE	-14.6680
FOUR	-20.1334
FIVE	-14.0825
SIX	-14.4977
SEVEN	-17.0890
EIGHT	-16.7658
NINE	-17.4949
TEN	-17.8979
YES	-21.6173
NO	-14.8297
SUCESSFULL	-19.3271
UNSUCESSFULL	-18.4982
% RECOGNITION	100%

TABLE III
PROBABILITY OF RECOGNIZING A WORD ONE AT 22 BITS/FRAME BY USING MSSVQ

NAME	PROBABILITY
ZERO	-6.3593
ONE	-13.1587
TWO	-119.6890
THREE	-15.3461
FOUR	-18.4924
FIVE	-19.3038
SIX	-20.5989
SEVEN	-18.4579
EIGHT	-15.3392
NINE	-13.8629
TEN	-15.4353
YES	-15.1308
NO	-16.9522
SUCESSFULL	-19.3051
UNSUCESSFULL	-19.5621
% RECOGNITION	93.33%

TABLE IV
PROBABILITY OF RECOGNIZING A WORD ONE AT 21 BITS/FRAME BY USING MSSVQ

NAME	PROBABILITY
ZERO	-12.7516
ONE	-16.0351
TWO	-18.8919
THREE	-19.7124
FOUR	-20.9550
FIVE	-18.4185
SIX	-19.7260
SEVEN	-17.8846
EIGHT	-18.4745
NINE	-20.5561
TEN	-19.9820
YES	-20.5743
NO	-17.5952
SUCESSFULL	-18.0211
UNSUCESSFULL	-19.3691
% RECOGNITION	93.33%

TABLE V
SPECTRAL DISTORTION, COMPLEXITY, AND MEMORY REQUIREMENTS FOR 3-PART SPLIT VECTOR QUANTIZATION TECHNIQUE

Bits / frame	SD(dB)	2-4 dB	>4dB	Complexity (kflops/frame)	ROM (floats)
24(8+8+8)	1.45	0.43	0	10.237	2560
23(7+8+8)	1.67	0.94	0	8.701	2176
22(7+7+8)	1.701	0.78	0.1	7.165	1792
21(7+7+7)	1.831	2.46	0.2	5.117	1280

TABLE VI
SPECTRAL DISTORTION, COMPLEXITY, AND MEMORY REQUIREMENTS FOR 3-STAGE MULTI STAGE VECTOR QUANTIZATION TECHNIQUE

Bits / frame	SD(dB)	2-4 dB	>4dB	Complexity (kflops/frame)	ROM (floats)
24(8+8+8)	0.984	1.38	0	30.717	7680
23(7+8+8)	1.238	1.2	0.1	25.597	6400
22(7+7+8)	1.345	0.85	0.13	20.477	5120
21(7+7+7)	1.4	1.08	0.3	15.357	3840

TABLE VII
SPECTRAL DISTORTION, COMPLEXITY, AND MEMORY REQUIREMENTS FOR 2-SWITCH 3-PART SWITCHED SPLIT VECTOR QUANTIZATION TECHNIQUE

Bits / frame	SD(dB)	2-4 dB	>4dB	Complexity (kflops/frame)	ROM (floats)
24(12+12)	0.957	1.06	0	8.78	4372
23(11+12)	1.113	1.29	0.14	7.244	3604
22(11+11)	1.119	0.52	1.3	5.196	2580
21(10+11)	1.127	1.3	0.56	4.428	2196

TABLE VIII
SPECTRAL DISTORTION, COMPLEXITY, AND MEMORY REQUIREMENTS FOR A 3-STAGE 2-SWITCH 3-PART MULTI SWITCHED SPLIT VECTOR QUANTIZATION

Bits / frame	SD(dB)	2-4 dB	>4dB	Complexity (kflops/frame)	ROM (floats)
24(8+8+8)	0.0322	0	0	0.9	396
23(7+8+8)	0.0381	0	0	0.836	364
22(7+7+8)	0.0373	0	0	0.772	332
21(7+7+7)	0.0377	0	0	0.708	300

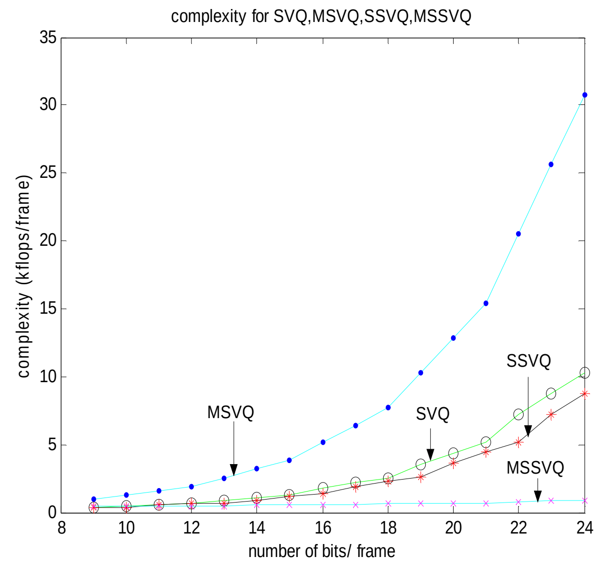

Fig. 3 Complexity for 3-part SVQ, 3-stage MSVQ, 2-switch 3-part SSVQ, and 3-stage 2-switch 3-part MSSVQ at various bit rates

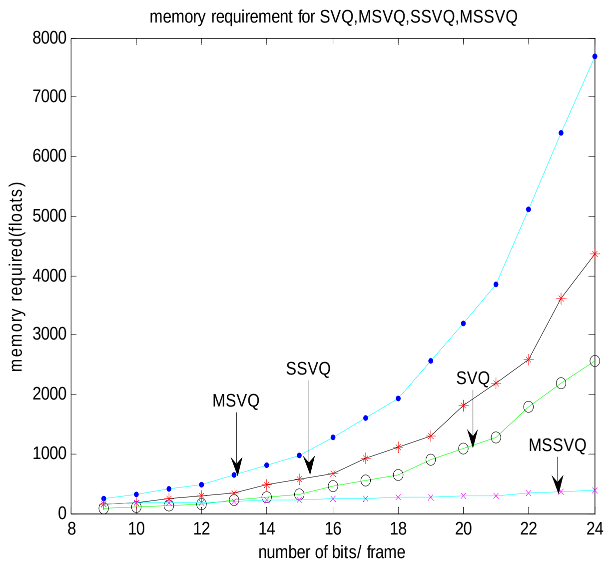

Fig. 4 Memory requirements for 3-part SVQ, 3-stage MSVQ, 2-switch 3-part SSVQ, and 3-stage 2-switch 3-part MSSVQ at various bit rates

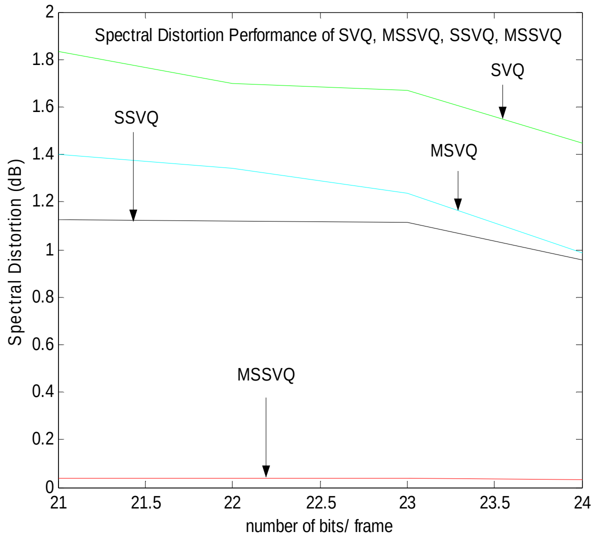

Fig. 5 Spectral Distortion Performance for 3-part SVQ, 3-stage MSVQ, 2-switch, 3-part SSVQ, and 3-stage, 2-switch, 3-part MSSVQ at various bit rates

VI. Conclusion

The Speech recognizer using HMM performs well for the coded output obtained by using MSSVQ, It has been observed that the percentage of recognition varies from 100% to 93.33% for different bit rates. Another advantage with MSSVQ is that it provides better trade-off between bit rate and spectral distortion performance, computational complexity, and memory requirements, when compared to other product code vector quantization schemes like Split vector quantization (SVQ), Multi stage vector quantization (MSVQ), and Switched Split vector quantization (SSVQ). So MSSVQ is proved to be better. When compared to all the product code vector quantization techniques. So MSSVQ is proved to be the better LPC coding technique for voice banking application. The performance can better improved by increasing the number of training vectors and bits for codebook generation, by increasing the number of states of an utterance, by using an efficient algorithm for the generation of emission matrix that takes into account the entire training set unless the K-means clustering that randomly picks vectors from the training set for the generation of an emission matrix., and by using a software having grater degree of precision. With Matlab it is difficult to obtain grater degree of precision when a large number of states are taken for a particular utterance.

Acknowledgment

The authors place on record their grateful thanks to the authorities of Chalapathi Institute of Technology, Mothadaka, Guntur, AP, INDIA, R.V.R & J.C.College of Engineering, Guntur, A.P, INDIA, K L College of Engineering, Guntur, A.P, INDIA, and Jawaharlal Nehru Technological University, College of Engineering, Hyderabad, INDIA for providing the facilities.

References

[1] Rabiner Lawrence, Juang Bing-Hwang, Fundamentals of speech Recognition, Prentice Hall, New Jersey, 1993, ISBN 0-13-015157-2.
[2] Lawrence R.Rabiner, A tutorial on Hidden Markov Models and selected applications in speech recognition, Poceedings of the IEEE, Vol 77, no.2, Feb 1989, pp.154-161.
[3] Rabiner L.R, Levinson S.E., Rosenberg A.E. & Wilpon J.G, Speaker independent recognition of isolated words using clustering techniques, IEEE Trans. Acoustics, Speech, Signal Proc., 1979, pp.336-349.
[4] M.Satya Sai Ram., P.Siddaiah., & M.MadhaviLatha, Multi Switched Split Vector Quantization of Narrow Band Speech Signals, Proceedings World Academy of Science, Engineering and Technology, WASET, Vol.27, Feb 2008, pp.236-239.
[5] M.Satya Sai Ram., P.Siddaiah., & M.MadhaviLatha, Multi Switched Split Vector Quantizer, International Journal of Computer, Information, and Systems science, and Engineering, IJCISSE, WASET, Vol.2, no.1, May 2008, pp.1-6.
[6] Paliwal. K.K, Atal. B.S, Efficient vector quantization of LPC Parameters at 24 bits/frame, IEEE Trans. Speech Audio Process, 1993, pp. 3-14.
[7] Stephen. So, & Paliwal. K. K, Efficient product code vector quantization using switched split vector quantizer, Digital Signal Processing journal, Elsevier, Vol 17, Jan 2007, pp.138-171.
[8] Bastiaan Kleijn. W, Tom Backstrom, & Paavo Alku, On Line Spectral Frequencies," IEEE Signal Processing Letters, Vol.10, no.3, 2003.
[9] Soong. F, Juang. B, Line spectrum pair (LSP) and speech data compression, IEEE Conference. On Acoustics, Speech Signal Processing, vol 9, no.1, Mar 1984, pp. 37-40.
[10] P. Kabal, & P. Rama Chandran, The Computation of Line Spectral Frequencies Using Chebyshev polynomials, IEEE Trans. On Acoustics, Speech Signal Processing, Vol 34, no.6, 1986, pp. 1419-1426.
[11] P. Lockwood and J. Boudy, .Experiments with a Nonlinear Spectral Subtraction (NSS), Hidden Markov Models and the Projection, for Robust Speech Recognition in Cars. Speech Communiaction, vol. 11, 1992 , pp. 215.228.
[12] S.F. Boll, Suppression of Acoustic Noise in Speech using Spectral Subtraction, IEEE Trans. on ASSP, vol. 27(2), 1979, pp.113-120.
[13] M. Berouti, R. Schwartz, and J. Makhoul, Enhancement of Speech Corrupted by Acoustic Noise. in Proc. ICASSP, 1979, pp. 208.211.
[14] Linde .Y, Buzo. A, & Gray. R.M, An Algorithm for Vector Quantizer Design, IEEE Trans.Commun, 28, Jan.1980, pp. 84-95.

M.Satya Sai Ram obtained B.Tech degree in Electronics and Communication Engineering from Nagarjuna University, Guntur in 2003. He received his M.Tech degree from Nagarjuna University, Guntur in 2005. He started his career as a lecturer at R.V.R & J.C. College of Engineering, Guntur, AP, INDIA in 2005 and promoted as a Sr.Lecturer in the year 2007. At present M.Satya Sai Ram is working as an Associate professor in the department of Electronics and Communication Engineering, at Chalapathi Institute of Technology, Mothadaka, Guntur, AP, INDIA. He actively involved in research and guiding Projects for Post Graduate students in the area of Speech & Signal Processing,. He has taught a wide variety of courses for UG students and guided several projects. He has published more than Six papers in International Conferences and Journals.

P. Siddaiah obtained B.Tech degree in Electronics and Communication Engineering from JNTU college of Engineering in 1988. He received his M.Tech degree from SV University, Tirupathi. He did his PhD program in JNTU, Hyderabad. He is the chief Investigator for several outsourcing project sponsored by Defense organizations and AICTE. He started his career as lecturer at SV University in 1993. At present Dr P. Siddaiah is working as an Professor & HOD in the department of Electronics and Communication Engineering, KL College of Engineering and actively involved in research and guiding students in the area of Antennas, Speech & Signal Processing,.. He has taught a wide variety of courses for UG & PG students and guided several projects. Several members pursuing their PhD degree under guidance. He has published several papers in National and International Journals and Conferences. He is the life member of FIETE, IE, and MISTE.

M. Madhavi Latha graduated in B. Tech from NU in 1986, Post Graduation in M.Tech from JNTU in 1993 and Ph. D from JNTU in 2002. She has been actively involved in research and guiding students in the area of Signal & Image Processing, VLSI (Mixed Signal design) and hardware implementation of Speech CODECs. She has published more than 30 papers in National/ International Conferences and Journals. Currently, she has been working as Professor in ECE, JNTU College of Engineering, Hyderabad, Andhra Pradesh. She is the life member of FIETE, MISTE, MIEEE.