VOWEL PHONEME RECOGNITION BASED ON AVERAGE ENERGY INFORMATION IN THE ZEROCROSSING INTERVALS AND ITS DISTRIBUTION USING ANN

Sunil Kumar R.K[1] and Lajish.V.L[2]

Department of Computer Science, University of Calicut, Kerala - 673635, INDIA
[1]sunilcsirc@rediffmail.com
[2]lajish@uoc.ac.in

ABSTRACT

Speech signal is modelled using the average energy of the signal in the zerocrossing intervals. Variation of these energies in the zerocrossing interval of the signal is studied and the distribution of this parameter through out the signal is evaluated. It is observed that the distribution patterns are similar for repeated utterances of the same vowels and varies from vowel to vowel. Credibility of the proposed parameter is verified over five Malayalam (one of the most popular Indian language) vowels using multilayer feed forward artificial neural network based recognition system. The performance of the system using additive white Gaussian noise corrupted speech is also studied for different SNR levels. From the experimental results it is evident that the average energy information in the zerocrossing intervals and its distributions can be effectively utilised for vowel phone classification and recognition.

KEYWORDS

Zerocrossing Intervals, Phoneme Recognition, ANN

1. INTRODUCTION

Parameterization of analog speech signal is the first step in the speech recognition process. Although various aspects of phoneme recognition have been investigated by researchers, the parameterization of analog speech and its computational complexity is still a headache for them. We are interested in a set of processing techniques used for phone recognition that are reasonably termed as time domain methods. By this we mean that the processing methods involve the wave form of the speech signal directly. Some examples of representation of speech signal in terms of time domain measurements include average zerocrossing rate, energy and autocorrelation function. The time domain method like zerocrossing analysis technique has been applied to several signal processing and signal analysis tasks. Some of this task include speech analysis and speech recognition [1] [2] [3] [4], electroencephalographic (EEG), biomedical applications, communication application, oceanographic analysis and many others [5]. The relative easy method by which Zerocrossing information can be extracted and its low cost implementation made this technique attractive. Zerocrossing rate is widely used in many speech analysis and recognition purposes. In application involving speech processing and speech recognition, additional interest in Zerocrossing analysis has gained impetus through observation of Licklider and Pollack who shoed that clipped speech is highly intelligible [6]. As a result, numerous speech analysis and recognition devices have been built utilizing Zerocrossing analysis techniques [7]. In

DOI : 10.5121/ijist.2012.2603

spite of this simplicity, the resulting representation provides a useful basis for estimating important features of the speech signal. Phoneme recognition has major contribution in designing practical efficient speech recognition systems. Many phoneme recognition systems are reported in literature [8] [9] [10] [11]. In this paper we modelled the speech signal with a computationally simple speech parameter using the average energy information in the zerocrossing intervals of the signal and used it for phoneme recognition applications. We have used five Malayalam (one of the most popular Indian language) vowels for the conduct of recognition experiments. The present work also illustrates how the proposed parameters can be effectively used in neural network based phoneme recognition systems. The paper is organized in five sections. Session 1.1 describes the representation of energy in the discrete speech signals. Session 2 describes the speech modelling technique using average energy in the zecrocrossing interval and the Session 3 describes the speech parameterization using the average energy in the zerocrossing interval distribution. Session 4 describes vowel recognition using Artificial Neural Network (ANN) and the final section deals the conclusion.

1.1. Energy of the Discrete Speech Signal

In a typical speech signal we can see that its certain properties considerably changes with time. For example, we can observe a significant variation in the peak amplitude of the signal and a considerable variation of fundamental frequency within voiced regions in a speech signal. These facts suggest that simple time domain processing techniques should be capable of providing useful information of signal features, such as intensity, excitation mode, pitch, and possibly even vocal tract parameters, such as formant frequencies. Most of the short time processing techniques that give time domain features (Q_n), can be mathematically represented as

$$Q_n = \sum_{m=-\infty}^{\infty} T[X(m)]W(n-m)$$

where T[] is the transformation matrix which may be either linear or nonlinear, X(m) represents the data sequence and W(n-m) represents a limited time window sequence. The energy of the discrete time signal is defined as

$$E = \sum_{m=-\infty}^{\infty} X^2(m)$$

Such a quantity has little meaning or utility for speech since it gives little information about time dependent properties of speech signal. We have observed that the amplitude of the speech signal varies appreciably with time. In particular, the amplitude of the unvoiced segment is generally much lower than amplitude of the voiced segment. The short time energy of the speech signal provides a convenient representation that reflects the amplitude variation and can be defined as

$$E_n = \sum_{m=-\infty}^{\infty} [X(m)\, W(n-m)]^2$$

The major significance of E_n is that it provides a basis for distinguishing voiced speech segment from unvoiced speech segment. It can be seen that the value of E_n for the unvoiced segments are significantly smaller than voiced segments. The energy function can also be used to locate approximately the time at which voiced speech become unvoiced speech and vice versa, and for high quality speech (high signal to noise ratio) the energy can be used to distinguish speech from

silence. The above discussion cites the importance of energy function (E_n) for speech analysis purpose.

2. SPEECH MODELLING USING AVERAGE ENERGY IN THE ZEROCROSSING INTERVAL

The speech production model [12] suggests that the energy of the voiced speech is concentrated below about 3 kHz, where as in the case of unvoiced speech, most of the energy is found at higher frequencies. Since high frequency implies high zerocrossing rate and low frequency implies low zerocrossing rate, there is strong correlation between zerocrossing rate and energy distribution with frequency [13] [14]. This motivates us to model the speech signal using average energy in zerocrossing interval of the signal. Consider the speech segment shown in Figure 1. The ZC^k_i shows the i^{th} zerocrossing and ZC^k_{i+1} shows the $i+1^{th}$ zerocrossing of k^{th} observation window. The time interval between these two points is called i^{th} zerocrossing interval T^k_i in the k^{th} observation window

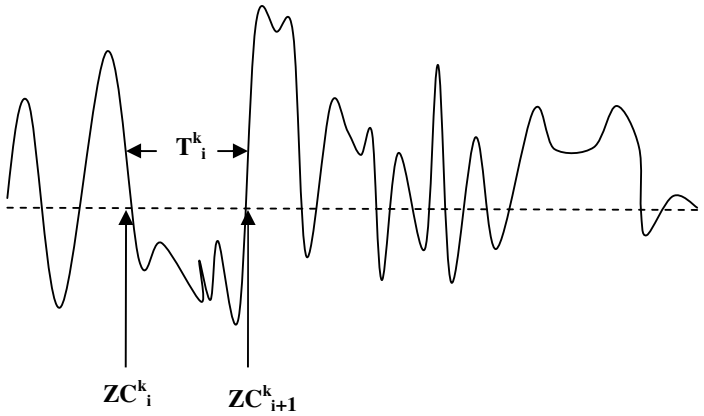

Figure 1. Speech segment in k^{th} observation window.

The average energy in the ith zerocrossing interval can be obtained by the expression

$$E_i^k = \frac{1}{T_i^k} \int_{ZC_i^k}^{ZC_{i+1}^k} X^2(t)\, dt$$

E^k_i is the average energy of the signal in T^k_i th zerocrossing interval of k^{th} observation window and $X(t)$ is the instantaneous signal amplitude. The aim of the present study is to find a robust coefficient for speech recognition application using the average energy in the zerocrossing interval (AEZI). An XY plot is generated by plotting index number of zero crossing interval along X axis and Average Energy in the Zerocrossing Interval (AEZI) along Y axis. Figure 2 (a-e) represents the average energy in the zerocrossing interval *vs* index number of the zerocrossing interval for the Malayalam vowels /a/, /i/, /u/, /e/, /o/ respectively.

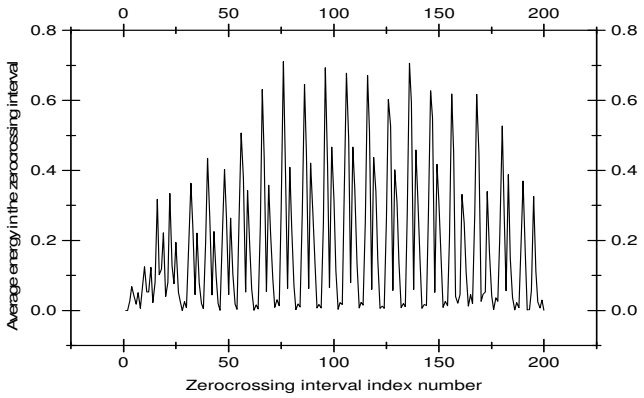

Figure 2 (a) Average energy in the zerocrossing interval vs zerocrossing interval index number of vowel /a/

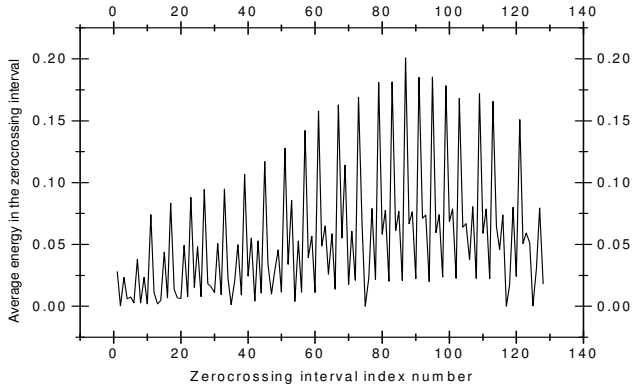

Figure 2 (b) Average energy in the zerocrossing interval vs zerocrossing interval index number of vowel /i/

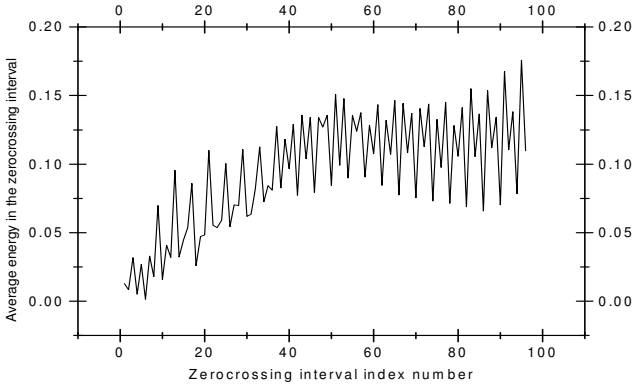

Figure 2 (c) Average energy in the zerocrossing interval vs zerocrossing interval index number of vowel /u/

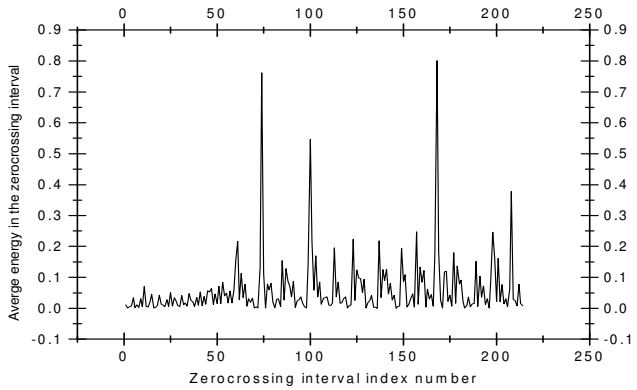

Figure 2 (d) Average energy in the zerocrossing interval vs zerocrossing interval index number of vowel /e/

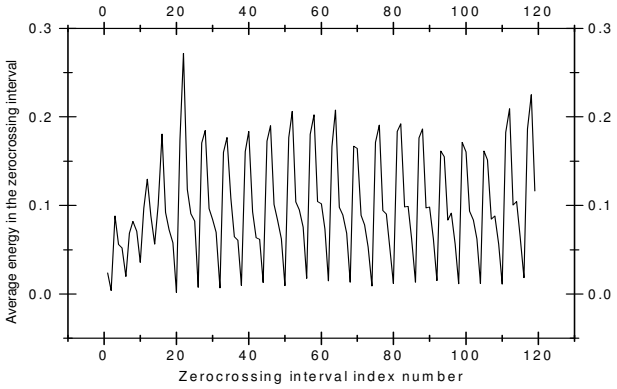

Figure 2 (e) Average energy in the zerocrossing interval vs zerocrossing interval index number of vowel /o/

From the Figures 2 (a-e), we can see that average energy in the consecutive zerocrossing interval are varying and its variation is different for different vowels . So this information can be used as a signature for identifying different vowels.

3. SPEECH MODELLING USING AVERAGE ENERGY IN THE ZEROCROSSING INTERVAL DISTRIBUTION

Let us define a range of energy values e_j = range [$E_{max}(j)$, $E_{min}(j)$], where $E_{max}(j)$ is the maximum and $E_{min}(j)$, the minimum of j^{th} range of energy. Now we define a parameter $Q^k(e_j)$ that gives the distribution of average energy in the zerocrossing interval of the signal $X(t)$ in a particular range of energy e_j of k^{th} observation window W^k. The function can be mathematically written as

$$Q^k(e_j) = \sum_{i=1}^{M^k} \eta\,(e_j, E_j^k)$$

where j = 1,2,……L ; k= 1,2,3 …….. N and $\eta(e_j, E^k_j) = 1$ if E^k_j lies in between the range specified for e_j and equal to zero otherwise.

For k = 1 we get the distribution values in all the L ranges as

$Q^1(e_1), Q^1(e_2), Q^1(e_3)$…….. $Q^1(e_L)$

For k = 2 the distribution values are

$Q^2(e_1), Q^2(e_2), Q^2(e_3)$…….. $Q^2(e_L)$

Or, in general, for $k = N$ the distribution values are

$Q^N(e_1), Q^N(e_2), Q^N(e_3)$…….. $Q^N(e_L)$

For j = 1 we get the total value of the distribution function as
$Q_{tot}(e_1) = Q^1(e_1)+ Q^2(e_1)+ Q^3(e_1)$……. $Q^N(e_1)$

For j =2,
$Q_{tot}(e_2) = Q^1(e_2)+ Q^2(e_2)+ Q^3(e_2)$……. $Q^N(e_2)$ etc.

Or, in general, for $j = L$ the total value of the distribution function will be

$Q_{tot}(e_L) = Q^1(e_L)+ Q^2(e_L)+ Q^3(e_L)$……. $Q^N(e_L)$

There fore,

$$Q_{tot}(e_j) = \sum_{k=1}^{N} Q^k(e_j);\ j = 1, 2, 3, \ldots, L$$

where $Q_{tot}(e_j)$ represents the distribution of average energy in the zerocrossing intervals in the range e_j.

3.1. Pattern formation

Speech signal is low pass filtered to 4kHz and sampled at 8kHz rate and digitized using 16 bit A/D converter. Five Malayalam vowels /a/, /i/, /u/, /e/, and /o/ uttered by a single speaker is used in the present study. From the band limited speech data, we have extracted Average Energy in the Zerocrossing Intervals (AEZI) using the equation

$$E_i^k = \frac{1}{T_i^k}\int_{ZC_i^k}^{ZC_{i+1}^k} X^2(t)\,dt$$

For finding the distribution of average energy in the zerocrossing interval for each vowel, first we fix a maximum threshold of energy value as 0.8 by examining the computed energy values of all the vowels. This upper threshold is divided into twenty uniform ranges (0-0.04 , 0.04-0.08,…. ……..0.76-0.8). The distribution of the average energy of the vowels in the zerocrossing interval among all the ranges mentioned above is obtained next. Figure 3(a-e) shows the distribution of average energy in the zerocrossing intervals of Malayalam vowels /a/, /i/, /u/, /e/ and /o/. The normal slim lines represent the distribution graph obtained using the repeated utterances of same vowels by the same speaker. The bold line shows the mean distribution plot.

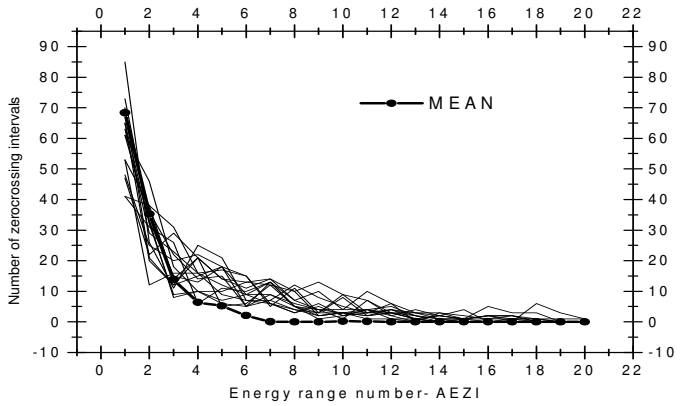

Figure 3 (a) Number of zerocrossing intervals *vs* Energy range number –AEZI of the vowel /a/

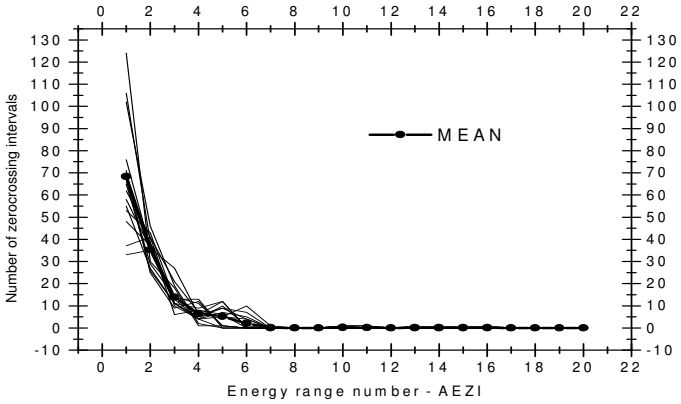

Figure 3 (b) Number of zerocrossing intervals *vs* Energy range number –AEZI of the vowel /i/

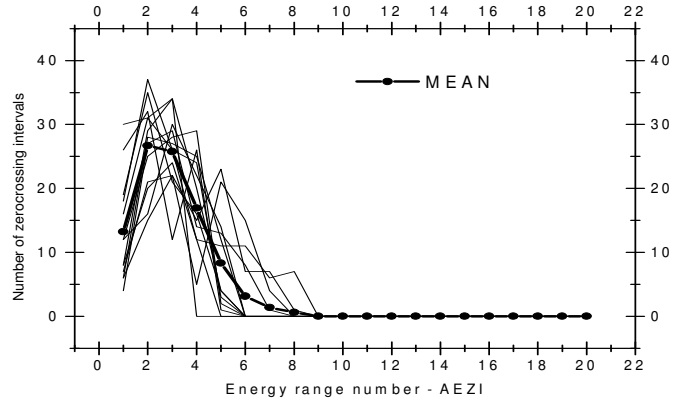

Figure 3 (c) Number of zerocrossing intervals *vs* Energy range number –AEZI of the vowel /u/

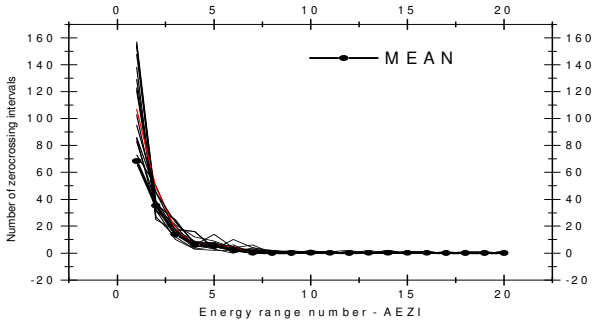

Figure 3 (d) Number of zerocrossing intervals *vs* Energy range number –AEZI of the vowel /e/

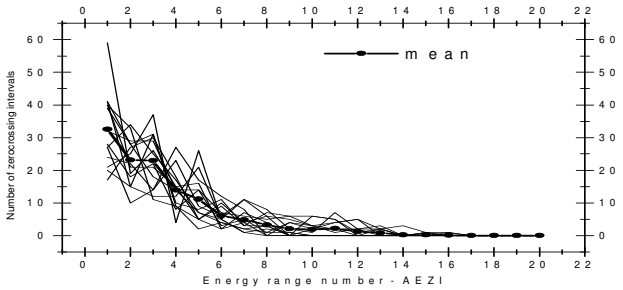

Figure 3 (e) Number of zerocrossing intervals *vs* Energy range number –AEZI of the vowel /o/

From the distribution graphs we can see that the distribution of average energy in the zerocrossing interval has almost similar pattern for repeated utterances of a particular vowel and varies from vowel to vowel. So this information can be used as a parameter for vowel recognition applications.

4. VOWEL RECOGNITION USING ANN

The application of artificial neural networks to speech recognition is the youngest and least understood of the recognition technologies. The ANN is based on the notion that complex "computing" operations can be implemented by massive integration of individual computing units, each of which performs an elementary computation. Artificial neural networks have several advantages relative to sequential machines. First, the ability to adapt is at the very center of ANN operations. Adaptation takes the form of adjusting the connection weights in order to achieve desired mappings. Furthermore ANN can continue to adapt and learn, which is extremely useful in processing and recognition of speech. Second, ANN tend to move robust or fault tolerant than Von Neumann machines because the network is composed of many interconnected neurons, all computing in parallel, and failure of a few processing units can often be compensated for by the redundancy in the network. Similarly ANN can often generalize from incomplete or noisy data.

Finally ANN when used as classifier does not require strong statistical characterization or parameterization of data [15] [16] [17] [18] [19]. These are the main motivations to choose artificial neural networks for phoneme recognition.

We used Multilayer Layer Feed Forward architecture for this experiment. The network consists of

20 input nodes and five output nodes, representing the five vowels. The network is trained using the energy information of the signal mentioned in Section 3. The experiment is repeated by changing the number of hidden layers and adding Additive White Gaussian Noise (AWGN) to the signal with different Signal to Noise Ratio (SNR). We used the database of 150 samples of each vowel spoken by a single male speaker for this purpose. A disjoint set of training and test patterns are formed by taking the 50% of the total patterns for each set. The error tolerance (Emax) is fixed as 0.001 and learning parameter (η) is chosen between 0.01 and 1.0. The training of the network is done using error back propagation learning method.

The recognition results are tabulated in Table 1 for neural network with three hidden layers. The recognition results indicates that the network shows poor recognition accuracy when the signal is added with Additive White Gaussian Noise of 0dB, 3dB and 10dB SNR. Above 20dB SNR, it shows better results. The average vowel recognition accuracy obtained for five Malayalam vowels in this experiment is 87.94%. This percentage is comparable with the phoneme recognition using spectral method based on ANN.

Table 1. Recognition accuracy for five Malayalam vowels with additive white Gaussian noise of different dB levels.

Vowel	Recognition accuracy %					
	0dB	3dB	10dB	20dB	30dB	Normal
/a/	33.3	33.3	46.6	66.6	86.6	86.6
/i/	26.6	33.3	40.0	53.3	86.6	86.6
/u/	13.3	26.6	33.3	73.3	86.6	86.6
/e/	13.3	33.3	40.0	60.0	93.3	93.3
/o/	26.6	33.3	46.6	73.3	86.6	86.6
Average	22.62	31.96	41.3	65.3	87.94	87.94

5. CONCLUSIONS

The speech signal is modelled using the average energy in the zerocrossing interval of the signal. Variation of these energies in the zerocrossing interval of the signal is studied and the distribution of this parameters through out the signal is evaluated. We found that the distribution patterns are almost similar for repeated utterances of the same vowels and varies from vowel to vowel. This distribution pattern is used for recognizing five Malayalam vowels using multilayer feed forward artificial neural network. The performance of this system using additive white Gaussian noise corrupted speech (vowel) is also studied for different SNR levels. It is found that recognition accuracy of the vowels is good when the signal is corrupted with low noise levels. The recognition accuracy obtained for five Malayalam vowels is 87.94 %. The present work can be extended, with the proposed neural network model and training algorithm, using vowels samples uttered by different male and female speakers under different age groups for analyzing the speaker independency of the proposed parameters.

REFERENCES

[1] Arai T & Yoshida Y. (1990) "Study on Zerocrossing of speech signals by means of analytic signal". Journal of Acoustical Society if Japan.; VolxxV+692, pp.242-246
Lee, S.hyun. & Kim Mi Na, (2008) "This is my paper", ABC *Transactions on ECE*, Vol. 10, No. 5, pp120-122.

[2] Russell J.Niederjohn,, Michal W. Krutz & Bruce M . Brown, (1987) "An experimental investigation of perceptual effects of altering the Zerocrossing of a speech signal", IEEE *Trans. Acoust. , speech and signal processing*, Vol. ASSP-35, No 5,pp.618-625.

[3] Erdol N, CastellucciaC, Zilouchian A, (1993) "Recovery of missing speech packet using the short time energy and Zerocrossing measurements", IEEE *Trans. Acoust. , speech and audio processing*, Vol.1.1, no3,pp.295-303.

[4] Sreenivas T.V and Russell J.Niederjohn, (1992), "Zerocrossing based spectral analysis and SVD spectral analysis for formant frequency estimation in noise", IEEE *trans. on signal processing*, Vol.40 No2.

[5] Russell J.Niederjohn, (1975), "A mathematical formulation and comparison of Zerocrossing Analysis Techniques which have been applied to automatic speech recognition" , IEEE *Trans. Accust. , speech and signal processing*, Vol. ASSP-23, No.4, pp373-380.

[6] Licklider J.C.R and Pollack I, (1948), "Effects of differentiation , Integration and infinite pack clipping upon intelligibility of of speech", J. *Accoust. Soc., Amer*, Vol-20, p42.

[7] Doh-Suk Kim, Jae Hoon Jeong, Jae Weon Kim & Soo Young Lee, (1996), "Feature extraction based on Zerocrossing with peak amplitudes for for robust speech recognition in noisy environments", IEEE *Trans. On Audio Electro acoust*., Vol AU 17, pp.61-64

[8] Chandrasekhar.C and Yegnanarayana .B, (1996), "Recognition of Stop – Consonant – Vowel (SCV) segments in continuous speech using neural network models", *Journal of Institution of Electronics and Telecommunication Engineers(IETE)*, Vol 42,pp.269-280

[9] Ki-Seok-Kim; Hee-Yeung-Hwang, (1991), "A study on the speech recognition of Korean phonemes using recurrent neural network models", *Transactions of the Korean Institute of Electrical Engineers*. Vol.40,no.8,p782-91.

[10] Rabiner L. R. & Juang.B.H, (1993), *Fundamentals of Speech Recognition* , Prentice Hall

[11] Sunilkumar R.K and Lajish.V.L, (2012), "Phone recognition using Zerocrossing Interval Distribution of Speech Patters and ANN", *International Journal of Speech Technology, (IJST)*, Springer, pp.1-7.

[12] Flangen J.L, (1972*), Speech Synthesis, analysis and perception*, New York, Springer Verlag

[13] Rabiner, L. R., and Schafer, R. W., (1978), *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ.

[14] L. R. Rabiner and M. R. Sambur, (1975), "An algorithm for determining the endpoints of isolated utterances", *Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315.

[15] Robert Hecht-Nielsen, (1990), *Neurocomputing*, Addison-Wesley.

[16] Sada Siva Varma A, Strube H.W, Rajesh Varma and Agarwal SS, (1996), "Recognition of Hindi Consonant Using Time Delay Neural Network", *Journal of the Acoustical Society Of India*, Vol.24,pp III-10.1-10.6.

[17] Richard P. Lippmann, (1989), "Review of Neural Networks for Speech Recognition", *Neural Computation*, Spring, Vol. 1, no. 1, pp. 1-38.

[18] W.Y. Huang and RP. Lippmann,(1987), "Neural Net and Traditional Classifiers," *Proc. IEEE Corif'. on Neural Information Processing Systems - Natural and Synthetic*, IEEE, New York.

[19] Simon King, Paul Taylor, (2000), "Detection of phonological features in continuous speech using neural networks", *Computer Speech & Language*, Elsevier, Vol.14, no.4, pp.333-353.

Authors

Dr Sunilkumar R.K earned his Ph.D in Speech Signal Processing from University of Calicut, Kerala, India in 2004. He has published several research papers in National and International levels in the area of signal processing and artificial neural networks. His research interest includes speech signal processing, neural networks and active noise cancellation.

Dr.Lajish.V.L has been associated with University of Calicut, Kerala, INDIA as Head of the Department of Computer Science. He has worked as Scientist R&D in TCS Innovation Labs, Tata Consultancy Services Ltd. Mumbai, prior to joining the University. His prime areas of research include Digital speech and image processing, Pattern recognition algorithms and Indian language script technology solutions for mobile devices. He has more than thirty research publications in journals and peer-reviewed National and International conferences to his credit. After his masters in Computer Science from Vellore Institute of Technology, Dr.Lajish earned his Ph.D in Computer Science from University of Calicut in 2007**.** He is a senior life member of International Association of Computer Science and Information Technology.