# EMOTIONAL TELUGU SPEECH SIGNALS CLASSIFICATION BASED ON K-NN CLASSIFIER

**P. Vijai Bhaskar[1], S. Ramamohana Rao[2]**

*[1]Professor, ECE Department, AVN Institute of Engg & Tech, A.P. India, [2] Principal, Geethanjali College of Engineering and Technology, A.P, India,* ***pvijaibhaskar@gmail.com, Principal.gcet@gmail.com***

## Abstract

*Speech processing is the study of speech signals, and the methods used to process them. In application such as speech coding, speech synthesis, speech recognition and speaker recognition technology, speech processing is employed. In speech classification, the computation of prosody effects from speech signals plays a major role. In emotional speech signals pitch and frequency is a most important parameters. Normally, the pitch value of sad and happy speech signals has a great difference and the frequency value of happy is higher than sad speech. But, in some cases the frequency of happy speech is nearly similar to sad speech or frequency of sad speech is similar to happy speech. In such situation, it is difficult to recognize the exact speech signal. To reduce such drawbacks, in this paper we propose a Telugu speech emotion classification system with three features like Energy Entropy, Short Time Energy, Zero Crossing Rate and K-NN classifier for the classification. Features are extracted from the speech signals and given to the K-NN. The implementation result shows the effectiveness of proposed speech emotion classification system in classifying the Telugu speech signals based on their prosody effects. The performance of the proposed speech emotion classification system is evaluated by conducting cross validation on the Telugu speech database.*

***Keywords:*** *Emotion Classification, K-NN classifier, Energy Entropy, Short Time Energy, Zero Crossing Rate.*

----------***----------

## 1. INTRODUCTION

Speech is the most desirable medium of communication between humans. There are several ways of characterizing the communications potential of speech. In general, speech coding can be considered to be a particular specialty in the broad field of speech processing, which also includes speech analysis and speech recognition. The purpose of speech coder is to convert an analogue speech signal into digital form for efficient transmission or storage and then to convert the received digital signal back to analogue [1]. Nowadays, speech technology has matured enough to be useful for many practical applications. Its good performance, however, depends on the availability of adequate training resources. There are many applications, where resources for the domain or language of interest are very limited. For example, in intelligence applications, it is often impossible to collect the necessary speech resources in advance as it is hard to predict which languages become the next ones of interest [4].

Currently, speech recognition applications are becoming increasingly advantageous. Also, different interactive speech aware applications are widely available in the market. In speech recognition, the sounds uttered by a speaker are converted to a sequence of words recognized by a listener. The logic of the process requires information to flow in one direction: from sounds to words. This direction of information flow is unavoidable and necessary for a speech recognition model to function [2]. In many practical situations, an automatic speech recognizer has to operate in several different but well-defined acoustic environments. For example, the same recognition task may be implemented using different microphones or transmission channels. In this situation, it may not be practical to recollect a speech corpus to train the acoustic models of the recognizer [3]. Speech input offers high bandwidth information and relative ease of use. It also permits the user's hands and eyes to be busy with a task, which is particularly valuable when users are in motion or in natural field settings [7]. Similarly, speech output is more impressive and understandable than the text output. Speech interfacing involves speech synthesis and speech recognition. Speech synthesizer takes the text as input and converts it into the speech output i.e. It acts as text to speech converter. Speech recognizer converts the spoken word into text [5].

Prosody refers to the supra segmental features of natural speech, such as rhythm and intonation [6]. Native speakers use prosody to convey paralinguistic information such as emphasis, intention, attitude and emotion. The prosody of a word sequence can be described by a set of prosodic variables such as prosodic phrase boundary, pitch accent, lexical stress, syllable position and hesitation, etc. Among these prosodic variables, pitch accent and into national phrase boundary have the most salient acoustic correlates and maybe most perceptually robust [8] [9]. Prosody is potentially convenient in automatic speech understanding systems for some reasons [10]. Prosody correlates with prosody may be used to

disambiguate syntactically distinct sentences with identical phoneme strings. Since prosody is ambiguous without phoneme information, and phonemes are ambiguous without prosodic information [11] [12]

The outline structure of the paper is organized as follows. In section 2, a brief review is made about the recent research works related to speech classification that is emphasized in relation to its problem statement. Next, Section 3 details the proposed speech emotion classification system and in Section 4 discusses about the implementation results and comparative results whereas as in Section 5 concludes the paper.

## 2. RELATED WORKS

A few of the most recent literature works in speech classification are reviewed below

Gyorgy Szaszaket al. [13] have developed a modeling acoustic processing approach for speech recognition. Acoustic processing and modeling of the supra-segmental characteristics of speech, with the aim of incorporating advanced syntactic and semantic level processing of spoken language for speech recognition/understanding tasks. The proposed modeling approach was similar to the one used in standard speech recognition, where basic HMM units were trained and were connected according to the dictionary and some grammar (language model) to obtain a recognition network, along which recognition could be interpreted also as an alignment process. In this proposed method, HMM framework was used to model speech prosody, and to perform initial syntactic and/or semantic level processing of the input speech in parallel for standard speech recognition. As acoustic–prosodic features, fundamental frequency and energy are used. For this, HMMs for the different types of word-stress unit contours were trained and then used for recognition and alignment of such units from the input speech. This method was also allows punctuation to be automatically marked.

Soheil Shafieea et al. [14] have proposed Speech Activity Detectors (SADs) technique for speech recognition tasks. In this proposed, a two-stage speech activity detection system was focused which at first take advantage of a voice activity detector to discard pause segments out of the audio signals; this was done even in presence of stationary background noises. To find the feature set in speech/non-speech classification, a large set of robust features was introduced; the optimal subset of these features was chosen by applying a Genetic Algorithm (GA) to the initial feature set. It had been discovered that Fractal dimensions of numeric series of prosodic features are the most speech/non-speech differentiating features. Models of the system were trained over a Farsi database, FARSDAT, however, the final experiments in the TIMIT English database had been conducted.

Raul Fernandez et al. [15] have proposed graphical models on the task of recognizing affective categories from prosody in both acted and natural speech. A strength of this approach was the integration and summarization of information using both local and global prosodic phenomena. In this framework speech was structurally modeled as a dynamically evolving hierarchical model in which levels of the hierarchy were determined by prosodic constituency and contain parameters that evolve according to dynamical systems. The acoustic parameters had been chosen to reflect four main components of the speech thought to reflect paralinguistic and affect-specific information: intonation, loudness, rhythm and voice quality. The model was then evaluated on two different corpora of fully spontaneous, effectively-colored, naturally occurring speech between people: Call Home English and BT Call Center. Here the ground truth labels were obtained from examining the agreement of 29 human coders labeling arousal and valence. Then finally detecting high arousal negative valence speech in call centers

Md. Mijanur Rahmanet al. [16] have proposed MFCC signal analysis technique for multi leveled pattern recognition task, which includes speech segmentation, classification, feature extraction and pattern recognition. In the proposed method, a blind speech segmentation procedure was used to segment the continuously spoken Bangla sentences into words/sub-words like units using the endpoint detection technique. These segmented words were classified according to the number of syllables and the sizes of the segmented words. The MFCC signal analysis technique was used to extract the features of speech words, which including windowing. The developed system achieved the segmentation accuracy rate at different from total 24 sub-classes of segmented words.

Abhishek Jaywant et al. [17] have characterized to category-specificity from speech prosody and facial expressions. In communication involved processing nonverbal emotional cues from auditory and visual stimuli. They employed a cross-modal priming task using emotional stimuli with the same valence but that diff erent in emotion category. After listening to angry, sad, disgusted, or neutral vocal primes, subjects rendered a facial aff ect decision about an emotionally congruent or incongruent face target. In this results revealed that participants made fewer errors when judging face targets that conveyed the same emotion as the vocal prime, and responded significantly faster for most emotions. Astonishingly, participants responded slower when the prime and target both conveyed disgust, perhaps due to attention biases for disgust-related stimuli. They find that vocal emotional expressions with similar valence are processed with category specificity and that discrete emotion knowledge implicitly aff ects the processing of emotional faces between sensory modalities.

## THE PROBLEM STATEMENT

Section 2 reviews the recent works related to the recognition or classification of speech signals based on the prosody effects. In continues speech signal classification, the prosody of continuous speech depends on many separate aspects, such as the meaning of the sentence and the speaker characteristics and emotions. In most of the research works, speech classification process based on prosody effects is done by using local and global features like, energy, pitch, linear predictive spectrum coding (LPCC), Mel-frequency spectrum coefficients (MFCC), and Mel-energy spectrum dynamic coefficients (MEDC), Intonation, Pitch contour, Vocal effort, etc. Normally, the pitch value of sad and happy speech signals has a great difference. Also, the frequency value of happy is higher than sad speech. But, in some cases the frequency of happy speech is nearly similar to sad speech or frequency of sad speech is similar to happy speech. In such situation, it is difficult to recognize the exact speech signal.

## 3. PROPOSED CLASSIFIER SYSTEM

. Our proposed emotion speech classification system classifies the speech signals based on their prosody effects by using K-NN Classifier. In our work, we have utilizes a Telugu speech signals to accomplish classification process. The proposed system mainly comprised of two stages namely, (i) Feature extraction (ii) Emotion classification. These two stages are consecutively performed and the more accurate results are occurred and are discussed in Section 3.1 and 3.2 respectively.

### 3.1 Feature Extraction

The Telugu speech database used in the proposed system consist speeches of seven male speakers and two female speakers with four emotions each. In feature extraction stage, there are three features are extracted from the input speech signals and given to the classification process. In speech classification feature extraction plays a most important role. Because the efficient features extraction from the input speech signals makes that the output to be more efficient and provide high classification accuracy. In our work, we extract three efficient features from the speech signals. The extracted three efficient features namely,

- Energy Entropy
- Short Time Energy
- Zero Crossing Rate

These features are extracted in our proposed system is explained below.

**(i) Energy Entropy ($E_e$)**
The energy level of an input speech words signal is rapidly changed and these sudden changes in the speech signals are measured. This measurement result is stated as energy entropy feature. To calculate this feature, the input speech word signals are divided into $f$ number of frames and calculate normalized energy for each frame. The energy entropy (Ee) feature is calculated by using the formula which is stated as follows,

$$E_e = -\sum_{k=0}^{f-1} \mu^2 . \log_2(\mu^2) \quad \text{-------- (1)}$$

$$\mu^2 = \sum_{b=1}^{N} \frac{\left( N * \frac{W_l}{S_b} \right)}{F} \quad \text{--------- (2)}$$

Where,

$\mu^2$ - is the normalized energy

$N$ - is the total number of blocks

$W_l$ - is the window length

$S_b$ - is the number of short blocks

$F$ - is the frequency

By exploiting the energy entropy equation is given in Equ. (1) Is applied to the speech word signals and obtain the energy entropy feature (Ee).

**(ii) Short Time Energy ($S_e$)**
The input speech signals energy level is to be increased suddenly. So we measure this energy increment level in speech signals is defined as short time energy. To calculate the short time energy the input signal is divided into $w$ number of windows and calculates the windowing function for each window. The short-time energy (Se) of speech signals replicates the amplitude variation and is described as follows,

$$S_e = \sum_{i=-\infty}^{\infty} x(i)^2 . h(w-i) \quad \text{------ (3)}$$

Where,

$x(i)$ - is the input signal

$h(w)$ - is the impulse response

By utilizing the equation is given in Equ. (3), the short time energy (Se) feature is calculated from the input speech word signals.

**(iii) Zero Crossing Rate (ZCR)**
Among these four features, the zero crossing rates is one of the most dominant features for speech signal processing. The zero crossing ratios are defined as the rate at which the speech signal crosses zero can provide information about the source

of its creation or the ratio of number of time domain zero crossings occurred to the frame length. The Zero crossing Rate (ZCR) is calculated by using the sign functions, which is stated below,

$$Z_{CR} = \frac{1}{M}\sum_{x=1}^{k-1}\frac{\text{sgn}\{y(x)\} - \text{sgn}\{y(x-1)\}}{2} \quad (4)$$

Where,

$M$ - is the length of the sample

$\text{sgn}\{y(x)\}$ - is the sign function

The sign function $\text{sgn}\{y(x)\}$ is defined as,

$$\text{sgn}\{y(x)\} = \begin{cases} 1; y(x) > 0 \\ 0; y(x) = 0 \\ -1; y(x) < 0 \end{cases} \quad (5)$$

The ZCR is calculated for each input speech word signals by using the Equ. (4).

### 3.2 Emotion Classification

In emotion classification the speech signals are classified by using the extracted features from the feature extraction stage. The extracted features from the speech signals are given to the K-NN classifier. The features are extracted for the training database speech signals and given to the K-NN classifier to perform the training process. During training stage, the speech signals corresponding three features are taken as input to the K-NN classifier. Here, we have taken three inputs as energy entropy (Ee), short-time energy (Se) and Zero crossing Rate (ZCR). The proposed emotion classification K-NN classifier Technique is mentioned below.

### 3.3 K-Nearest Neighbor Technique as an Emotion Recognizer

A more general version of the nearest neighbor technique bases the classification of an unknown sample on the "votes" of K of its nearest neighbor rather than on only it's on single nearest neighbor. The K-Nearest Neighbor classification procedure is denoted is denoted by K-NN. If the costs of error are equal for each class, the estimated class of an unknown sample is chosen to be the class that is most commonly represented in the collection of its K nearest neighbors. Among the various methods of supervised statistical pattern recognition, the Nearest Neighbor is the most traditional one, it does not consider a priori assumptions about the distributions from which the training examples are drawn. It involves a training set of all cases. A new sample is classified by calculating the distance to the nearest training case, the sign of that point then determines the classification of the sample. The K-NN classifier extends this idea by taking the K nearest points and assigning the sign of the majority. It is common to select K small and odd to break ties (typically 1, 3 or 5). Larger K values help reduce the effects of noisy points within the training data set, and the choice of K is often performed through cross validation. In this way, given a input test sample vector of features x of dimension n, we estimate its Euclidean distance d equation 1 with all the training samples (y) and classify to the class of the minimal distance.

$$q(x,y) = \sqrt{\sum_{i=1}^{n} (x_i^2 - y_i^2)} \text{--------(6)}$$

The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the classification phase, K is a user-defined constant, and an unlabelled vector (a query or test point) is classified by assigning the label which is most frequent among the K training samples nearest to that query point. Usually Euclidean distance is used as the distance metric, however this is only applicable to continuous variables.

### 3.4 K-NN Algorithm:

The k-NN algorithm can also be adapted for use in estimating continuous variables. One such implementation uses an inverse distance weighted average of the k-nearest multivariate neighbors. This algorithm functions as follows: Compute Euclidean or Mahalanobis distance from target plot to those that were sampled.

1. Order samples taking for account and calculate distances.
2. Choose heuristically optimal K nearest neighbor based on root mean square error q(x, y) done by cross validation technique.
3. Calculate an inverse distance weighted average with the k-nearest multivariate neighbors.

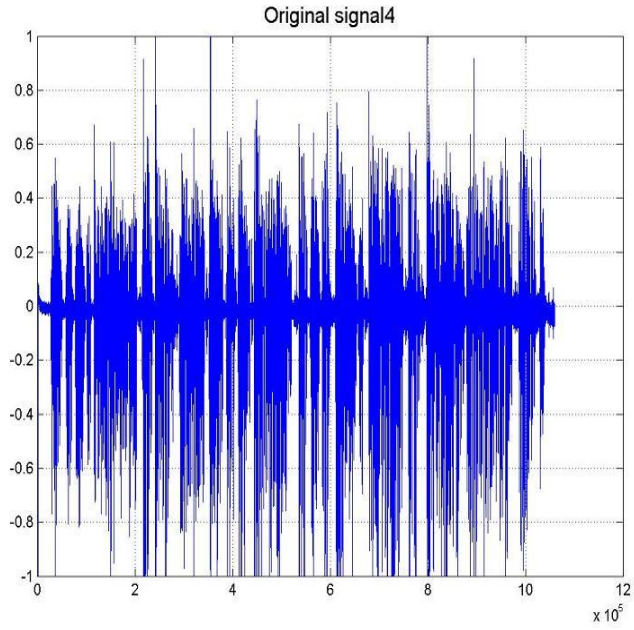
## 4. EXPERIMENTAL RESULTS

The proposed Telugu speech emotion classification technique is implemented in the working platform of MATLAB (version 7.12) with machine configuration as follows

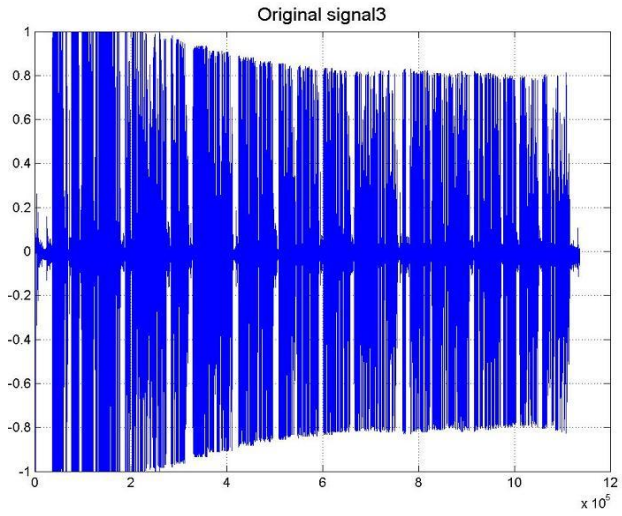
Processor: Intel core i5
OS: Windows xp
CPU speed: 3.20 GHz, RAM: 4GB

The performance of the proposed system is evaluated with different person's emotion speech signals and the results are compared against the existing techniques. The input speech signals are classified using the proposed speech classification

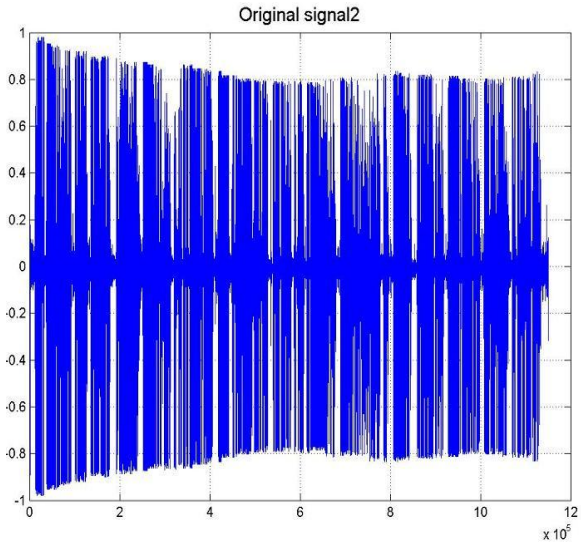
system using K-NN classifier. The sample (i) input normal, happy, sad and angry emotion speech signals (ii) Extracted features zero crossing rate, energy entropy, short time energy for normal, happy, sad and angry speech signals are shown in the fig.1



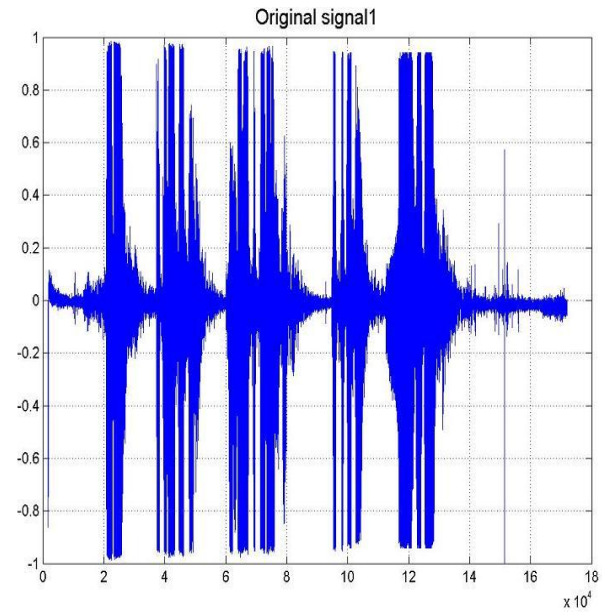

**(a)**




**(b)**

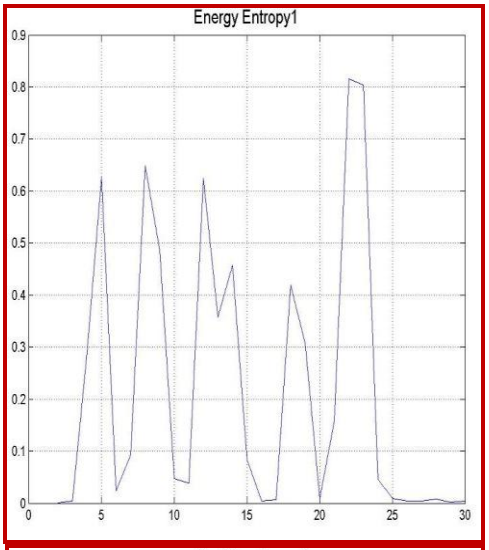


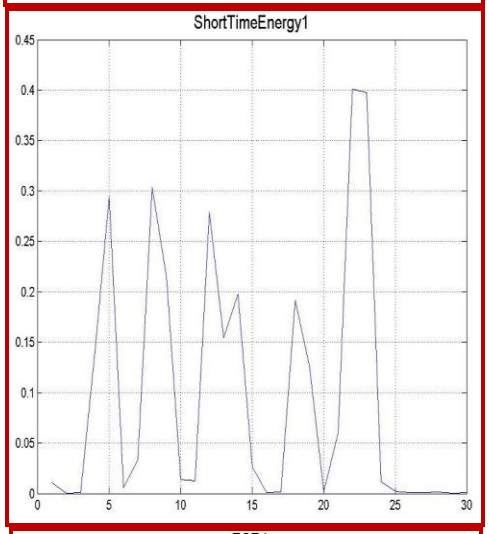

**(c)**

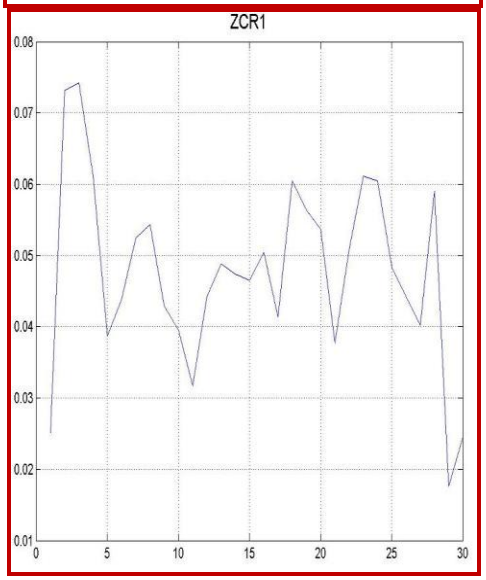



**(d)**

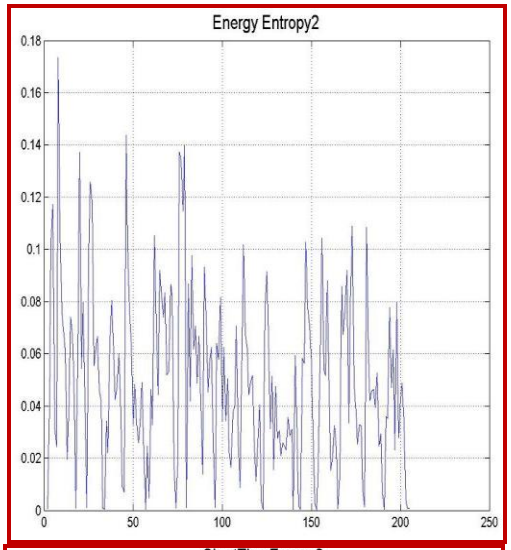
**1(i)**

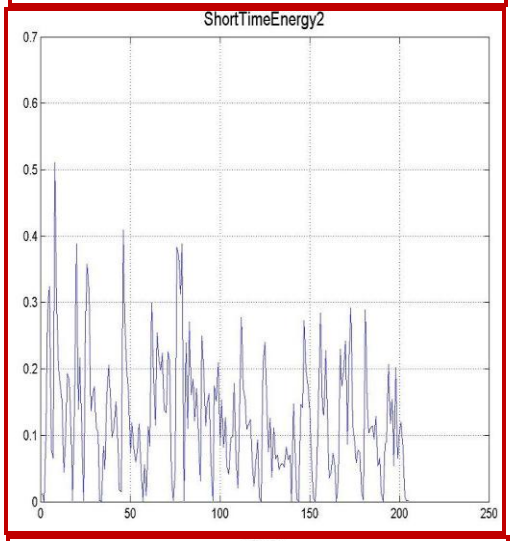

Energy Entropy1

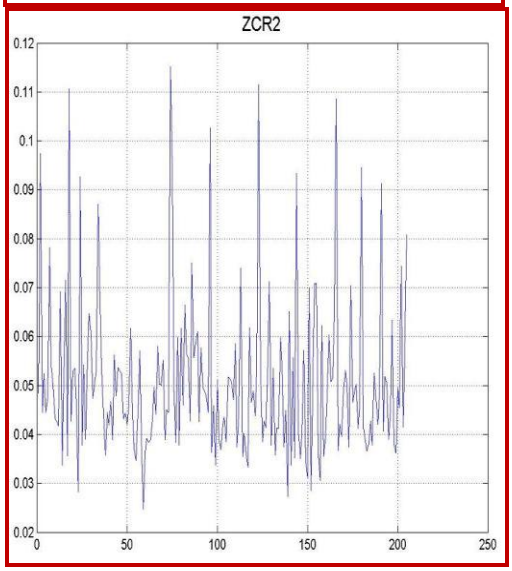

ShortTimeEnergy1


ZCR1

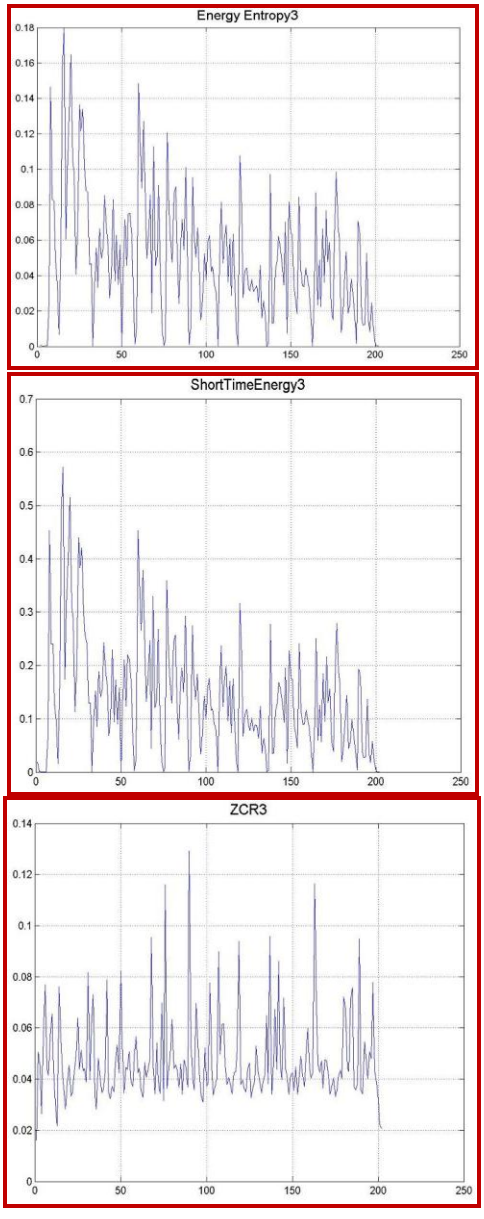
**(a)**


Energy Entropy2

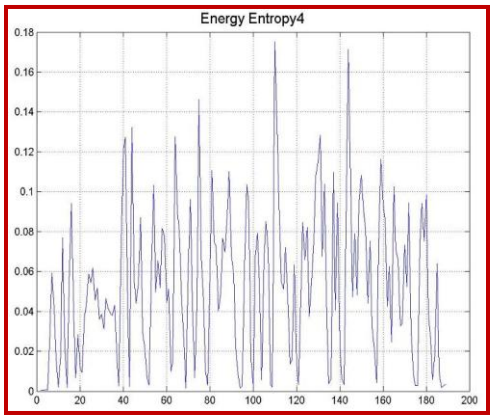

ShortTimeEnergy2

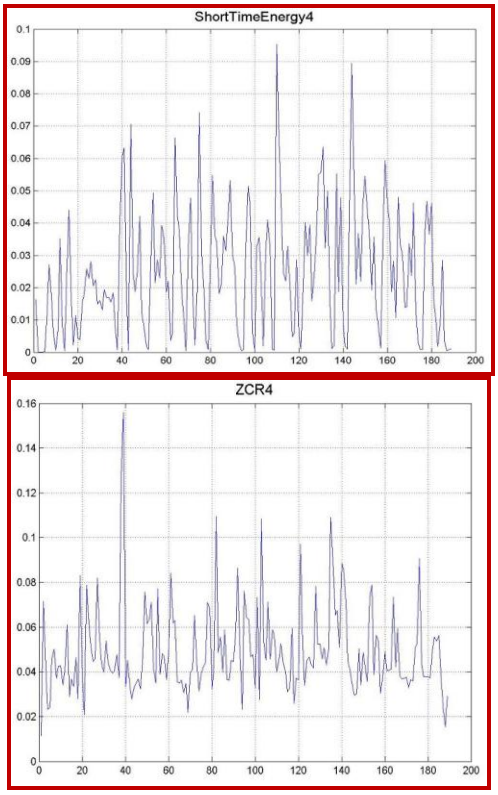

ZCR2

**(b)**




**(c)**




**(d)**

**1(ii)**

**Figure 1:** (i) Input (a) angry, (b) happy, (c) normal and (d) sad emotion speech signals (ii) Extracted features of (a) angry, (b) happy, (c) normal and (d) sad signals

The performance of proposed speech classification system is analyzed by using the statistical measures. The statistical measures [18] are used to measure the speech classification performance. The performance analyses have shown that the proposed system has been successfully classifies the input speech signals based on their prosody which the speech signals belongs to.

The Telugu speech signal database is created with 9 persons, each person has four emotional speech signals are normal, happy, sad and angry. These 9 person's emotion speech signals are given to the one cross validation process and the performance measures of the classification results are listed in Table .1

The K-NN classifier performance measures for the same Telugu speech database are given in Table .1.

**Table 1:** Speech emotion classification Statistical performance measures of K-NN for 9 different Speakers.

| Experiments | Sensitivity | FPR | Accuracy | Specificity | PPR | NPV | FPR | MCC |
|---|---|---|---|---|---|---|---|---|
| 1 | 25 | 75 | 25 | 25 | 25 | 25 | 75 | 13 |
| 2 | 0 | 67 | 25 | 33 | 0 | 33 | 100 | -12 |
| 3 | 25 | 75 | 25 | 25 | 13 | 38 | 88 | -9 |
| 4 | 25 | 75 | 25 | 25 | 25 | 25 | 75 | 13 |
| 5 | 25 | 75 | 25 | 25 | 25 | 25 | 75 | 13 |
| 6 | 25 | 75 | 25 | 25 | 25 | 25 | 75 | 13 |
| 7 | 25 | 75 | 25 | 25 | 25 | 25 | 75 | 13 |
| 8 | 25 | 75 | 25 | 25 | 25 | 25 | 75 | 13 |
| 9 | 25 | 75 | 25 | 25 | 25 | 25 | 75 | 13 |

As can be seen from table 1, our proposed emotion classification K-NN system has given a considerable and moderate classification performance.

## CONCLUSIONS

In this paper, we proposed a Telugu speech emotion classification system using K-NN Classifier. Here, the classification process is made by extracting three features like entropy, short time energy and zero crossing rates from the input speech signals. The computed features were given to the K-NN to Classifier accomplish the training process. The proposed speech emotion classification system classifies the Telugu speech signals based on their prosody by using the K-NN classifier. During testing, if a set of emotional speech signals is given as input it classifies the speech signals based on the prosody which the speech signal belongs to. Human emotions can be recognized from speech signals when facial expressions or biological signals are not available. In this work Emotions are recognized from speech signals using real time database. In this work we presented an approach to emotion recognition from speech signal. The future work will be to conduct comparative study of various classifier using different parameter selection method to improve and accuracy.

## REFERENCES

[1] Marwan Al-Akaidi, "Introduction to speech processing", Fractal Speech Processing-Cambridge University Press, 2012

[2] Dennis Norris, James M. McQueen and Anne Cutler, "Merging information in speech recognition: Feedback is Never necessary", Behavioral and Brain Sciences, Vol. 23, pp. 299-370, 2000

[3] Leonardo Neumeyer and Mitchel Weintraub, "Probabilistic Optimum Filtering for Robust Speech Recognition", In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Adelaide, SA, Australia, Vol. 1, pp. I/417 - I/420, 1994

[4] Lukas Burget, Petr Schwarz, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondrej Glembek, Nagendra Goel, Martin Karafiat, Daniel Povey, Ariya Rastrow, Richard C. Rose, Samuel Thomas, "Multilingual Acoustic Modeling for Speech Recognition Based on Subspace Gaussian Mixture Models", In Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing, Dallas, TX, pp. 4334 - 4337, 2010,

[5] Kuldeep Kumar and R. K. Aggarwal, "Hindi Speech Recognition System Using HTK", International Journal of Computing and Business Research, Vol. 2, No. 2, May 2011

[6]. Ken Chen, Mark Hasegawa-Johnson and Aaron Cohen, "An Automatic Prosody Labeling System Using Ann-Based Syntactic-Prosodic Model and GMM-Based Acoustic-Prosodic Model", IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 509-512, 2004.

[7] Vimala and Radha, "A Review on Speech Recognition Challenges and Approaches", World of Computer Science and Information Technology Journal, Vol. 2, No. 1, pp. 1-7, 2012

[8] Omair Khan, Wasfi G. Al-Khatib, Cheded Lahouari, "Detection of Questions in Arabic Audio Monologues Using Prosodic Features", Ninth IEEE International Symposium on Multimedia, pp. 29- 36, 2007

[9]. Stefanie Shattuck-Hufnage and Alice E. Turk, "A Prosody Tutorial for Investigators of Auditory Sentence

Processing", Journal of Psycholinguistic Research, Vol. 25, No. 2, pp. 193-247, 1996
[10] Elizabeth Shriberga, Andreas Stolckea, Dilek Hakkani-Türb, Gökhan Tur, "Prosody-based automatic segmentation of speech into sentences and topics", Speech Communication, Vol. 32, No. 1–2, pp. 127–154, 2000
[11] Ken Chen, Mark Hasegawa-Johnson, Aaron Cohen, Sarah Borys, Sung-Suk Kim, Jennifer Cole, and Jeung-Yoon Choi, "Prosody Dependent Speech Recognition on Radio News Corpus of American English", IEEE Transactions on Speech and Audio Processing, Vol. 13, No. 6, pp. 1-15, 2005
[12] Mark Hasegawa-Johnson, Ken Chen, Jennifer Cole, Sarah Borys, Sung-Suk Kim, Aaron Cohen, Tong Zhang, Jeung-Yoon Choi, Heejin Kim, Taejin Yoon, Sandra Chavarria, "Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus", Speech Communication, Vol. 46, No. 3–4, pp. 418–439, 2005
[13] KlaraVicsi, Gyorgy Szaszak, "Using prosody to improve automatic speech recognition", Speech Communication, Vol. 52, No. 5, pp. 413–426, 2010
[14] Soheil Shafieea, Farshad Almasganja, Bahram Vazirnezhada, Ayyoob Jafari, "A two-stage speech activity detection system considering fractal aspects of prosody", Pattern Recognition Letters, Vol. 31, No. 9, pp. 936–948, 2010
[15] Raul Fernandez, Rosalind Picard, "Recognizing affect from speech prosody using hierarchical graphical models", Journal Speech Communication, Vol. 53, No. 9-10, pp. 1088-1103, 2011
[16] Md. Mijanur Rahman, Md. Farukuzzaman Khan and Md. Al-Amin Bhuiyan, "Continuous Bangla Speech Segmentation, Classification and Feature Extraction", IJCSI International Journal of Computer Science Issues, Vol. 9, No. 2, No 1, pp. 67-75, 2012
[17] Abhishek Jaywant, Marc D. Pell, "Categorical processing of negative emotions from speech prosody", Speech Communication, Vol. 54, No. 1, pp. 1–10, 2012
[18]http://en.wikipedia.org/wiki/Sensitivity_and_specificity

## BIOGRAPHIES:

**P. Vijay Bhaskar,** Head of the ECE Department, AVN Institute of Engg.&Tech. –HYDERABAD obtained his B. Tech., from JNTU, Kakinada and M. Tech., from JNTU, Hyderabad, and both First Class with Distinction and pursuing PhD. from JNTU, Hyderabad, and also having the overall teaching experience of 18 Years and Published 09 papers in various National and International conferences, and guided good number of B. Tech., and M. Tech., Projects

**S. Rama Mohana Rao**, who served for 25 years in ISRO Vikram Sarabhai Space Centre Trivandrum from 1971-1996 in various capabilities such as Head, ELP, PCF, EFF and the last being as Deputy Project Director ESP. He has also served for 15 years in various educational institutions as Professor, HOD & Principal. who is an eminent personality known for his versatility, and obtained his Ph D from IISC, Bangalore, and also his credit publications in International & National journals. He has over15 years of experience as Principal at reputed Engineering colleges.