International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Performance Improvement in the Analysis and Classification of Telugu Emotion Speech Signals Based on FFBNN

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

[1]AVN Institute of Engineering and Technology, Hyderabad, India

[2]Former Principal, Geethanjali College of Engg.& Technology, Hyderabad, India

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 and use neural network for the classification. Features are extracted with optimal window size from the speech signals and given to the FFBNN. The well trained FFBNN is tested with more number of speech signals with prosody effects. 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 change the window size while extracting the features.***

Keywords: Emotion Classification, Feed Forward Back Propagation Neural Network (FFBNN), 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. There are many applications, where resources for the domain or language of interest are very limited.

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.

Prosody refers to the suprasegmental 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 [9] [10]. Prosody is potentially convenient in automatic speech understanding systems for some reasons [12]. 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 [12] [13]. The outline structure of the paper is organized as follows. Section 2 details the problem statement and section 3 discusses the proposed speech emotion classification system and about the implementation results and comparative results whereas as in Section 4 concludes the paper.

2. Objective of the Paper

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 Emotion Speech Classification System on Telugu Speech Signals

Our proposed emotion speech classification system classifies the speech signals based on their prosody effects by using FFBNN. 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

Licensed Under Creative Commons Attribution CC BY

occurred and are discussed in Section 3.1 and 3.2 respectively.

3.1 Feature Extraction

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 (E_e) feature is calculated by using the formula which is stated as follows,

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

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

Where,

μ^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 (E_e).

(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 (S_e) of speech signals replicates the amplitude variation and is described as follows,

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

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 (S_e) 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,

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

Where,

M - is the length of the sample

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

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

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

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 FFBNN. The features are extracted for the training database speech signals and given to the FFBNN to perform the training process. During training stage, the speech signals corresponding three features are taken as input to the FFBNN. Here, we have taken three inputs nodes as energy entropy (E_e), short-time energy (S_e) and Zero crossing Rate (Z_{CR}), N_d number of hidden layers and one output layer, which is a prosody effect of the given input signal. The proposed emotion classification FFBNN structure is shown in Fig. 1.

Licensed Under Creative Commons Attribution CC BY

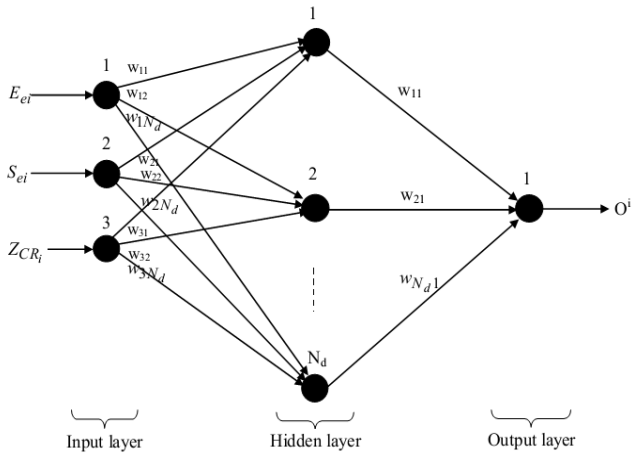

Figure 1: Structure of FFBNN in Emotion Classification

Initially, the input feature values are transmitted to the hidden layer and then, to the output layer. Each node in the hidden layer gets input from the input layer, which are multiplexed with suitable weights and summed. The hidden layer input value calculation function is called as bias function, which is described below,

$$H_i = \beta + \sum_{h=1}^{N_d} (w_h E_{eih} + w_h S_{eih} + w_h Z_{CRih}) \quad (6)$$

In Eq. (6), E_{ei}, S_{ei} and Z_{CRi} are the feature values of the i[th] person speech signal. The activation function in the output layer is given in Equ. (7). The output values from the output layer are compared with target values and the learning error rate for the neural network is computed, which is given in equation (8).

$$\alpha = \frac{1}{1+e^{-H_i}} \quad (7)$$

$$\lambda = \frac{1}{N_d} \sum_{h=0}^{N_d - 1} D_h^i - O_h^i \quad (8)$$

In eqn. (8),

λ - is the learning error rate

D_h^i - is the desired output

O_h^i - is the actual output.

The error between the nodes is transmitted back to the hidden layer and this process is called the backward pass of the back propagation algorithm. The reduction of error by back propagation algorithm is described in the subsequent steps.

(i) Initially, the weights are assigned to hidden layer neurons. The input layer has a constant weight, whereas the weights for output layer neurons are chosen arbitrarily. Subsequently, the bias function and output layer activation function are computed by using the Eq. (6) and (7).

(ii) Next, the back propagation error is computed for each node and the weights are updated by using the Eqn. (9).

$$w_{ih} = w_{ih} + \Delta w_{ih} \quad (9)$$

Where, the weight Δw_{ih} is changed, which is given as,

$$\Delta w_{ih} = \eta . N_{ih} . E^{(\varphi)} \quad (10)$$

Where, η is the learning rate that normally ranges from 0.2 to 0.5, and $E^{(\varphi)}$ is the BP error. The bias function, activation function, and BP error calculation process are continued till the BP error gets reduced i.e., $E^{(\varphi)} < 0.1$. If the BP error reaches a minimum value, then the FFBNN is well trained by the speech signals feature values for performing the emotion classification. The well trained FFBNN provides an accurate classification results for the input emotion speech signals. During testing, the features are extracted from the testing database speech signals and these extracted features are given to the trained FFBNN. The well trained FFBNN classify the input speech signals based on the prosody which the speech signal belongs to.

4. Experimental Results

The proposed Telugu speech emotion classification system 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 FFBNN. As an example the sample (i) input for sad emotion speech signal shown in fig.1 (ii) Extracted features zero crossing rate, energy entropy, short time energy for angry speech signal are shown in the fig.21,2.2 and 2.3.and fig.2.4 represents neural network training and fig. 2.5 represents classifier output.

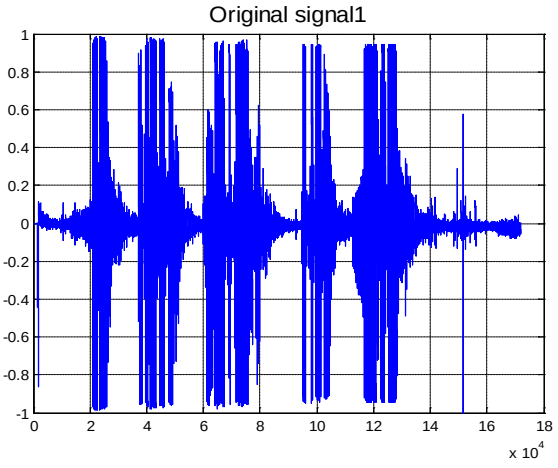

Figure 2.1

Licensed Under Creative Commons Attribution CC BY

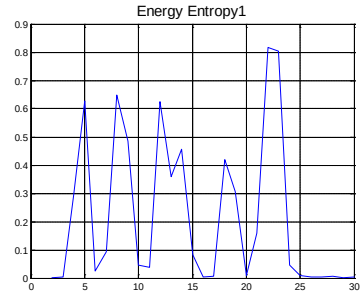

Figure 2 .1

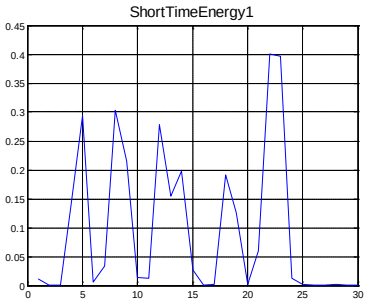

Figure 2.2

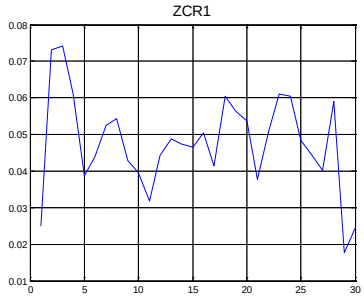

Figure 2.3

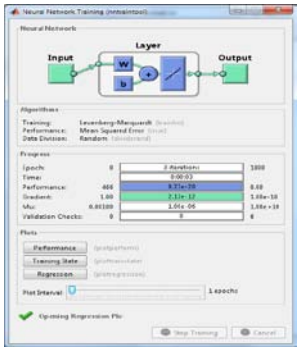

Figure 2.4

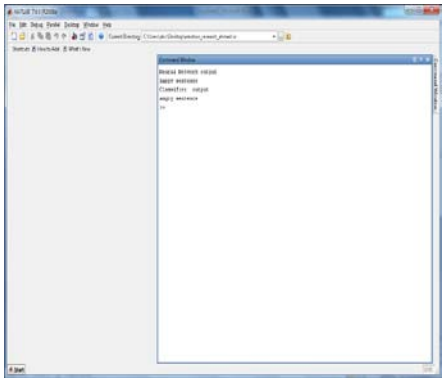

Figure 2.5

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 9 experiments average accuracy, sensitivity, and specificity values of proposed FFBNN are 76%, 47%, and 86% respectively. But, the K-NN classifier has given low mean statistical measures. 25%, 23% and 26%.The high accuracy results show that our proposed FFBNN system classifies the Emotion speech signals accurately. From the above it is clear that the sensitivity, specificity, and accuracy values of our proposed classification system are high compared to the K-NN classifier. Hence, our proposed Telugu speech emotion classification system based on FFBNN has accurately classifies the speech signals based on the emotion which the speech signal belongs to.

5. Conclusion

In this paper, we proposed a Telugu speech emotion classification system using FFBNN by incorporating suitable window size during the extraction of features. Here, the classification process is made by extracting three features like entropy, short time energy and zero crossing rates from the input speech signals by adjusting the window suitably to its threshold value. The computed features were given to the FFBNN to accomplish the training process. The proposed speech emotion classification system classifies the Telugu speech signals based on their prosody by using the FFBNN. 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. The performance of our proposed speech emotion classification system is compared with K-NN classifier. The comparison results show that our proposed Telugu speech emotion classification system using FFBNN is effective and depends on the size of the window that is employed for feature extraction in emotional Telugu speech classification.

References

[1] P.Vijai Bhaskar, S. Ramamohana Rao, M.Mahesh, "Analysis and Classification of Telugu Emotion Speech Signals based on FFBNN", Elsevier, ICACM-2013,GRIET,ISBN No: 9789351071495 Page Nos. 331-337

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

[3] 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

[4] 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

[5] 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,

Licensed Under Creative Commons Attribution CC BY

"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,

[6] 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

[7] 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.

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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

[14] KlaraVicsi, Gyorgy Szaszak, "Using prosody to improve automatic speech recognition", Speech Communication, Vol. 52, No. 5, pp. 413–426, 2010

[15] 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

[16] 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

[17] 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

[18] Abhishek Jaywant, Marc D. Pell, "Categorical processing of negative emotions from speech prosody", Speech Communication, Vol. 54, No. 1, pp. 1–10, 2012

Author Profile

Mr P. Vijai Bhaskar, Head of the ECE Department-AVNIET –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 08 papers in various National and International conferences, and guided good number of B. Tech., and M. Tech., Projects

Prof. Dr. 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 over 15 years of experience as Principal at reputed Engineering colleges.

Licensed Under Creative Commons Attribution CC BY