Chapter-5
Analysis of Emotion recognition system for Telugu speech

5.1. Introduction

Emotions accompany everywhere and at every moment in the life of human beings. The different types of emotions for a human being are Happy, Sad, Bore, Anger, Disgust, Fear and Neutral. Emotions are communicated through speech signals that constitute 38% of the whole communicated speech [105]. Hence speech emotion recognition (SER) becomes important which automatically classifies the emotional state of a speaker from speech signals into one of the several basic emotions.

Humans express their intentions and feelings by emotions. To convey the message correctly, the interface for machines and human beings, must understand the emotions properly. For better interaction between the machine and humans, the state of the emotion of the speaker must be adapted correctly by the machine. This will help the machine to reply back to humans with appropriate emotion.

In this section the methodology, data acquisition, feature extraction process, results and discussions, along with conclusions of the proposed emotion recognition system are discussed.

5.2. Methodology and block diagram

The three important steps involved in any pattern recognition problem are database collection, features extraction and classification. Telugu speech emotion recognition system requires

emotion database for both training and testing samples. As standard Telugu speech database is not available, a database is developed in the speech laboratory. After the database is developed, the features are extracted for both the training and testing samples.

A classification technique, namely Nearest Neighborhood classifier (NNC) is employed and the classification algorithm is trained with the feature vectors of training speech samples. Then the emotion in the test speech sample is recognized. The basic methodology of emotion recognition system is shown in fig 5.1.

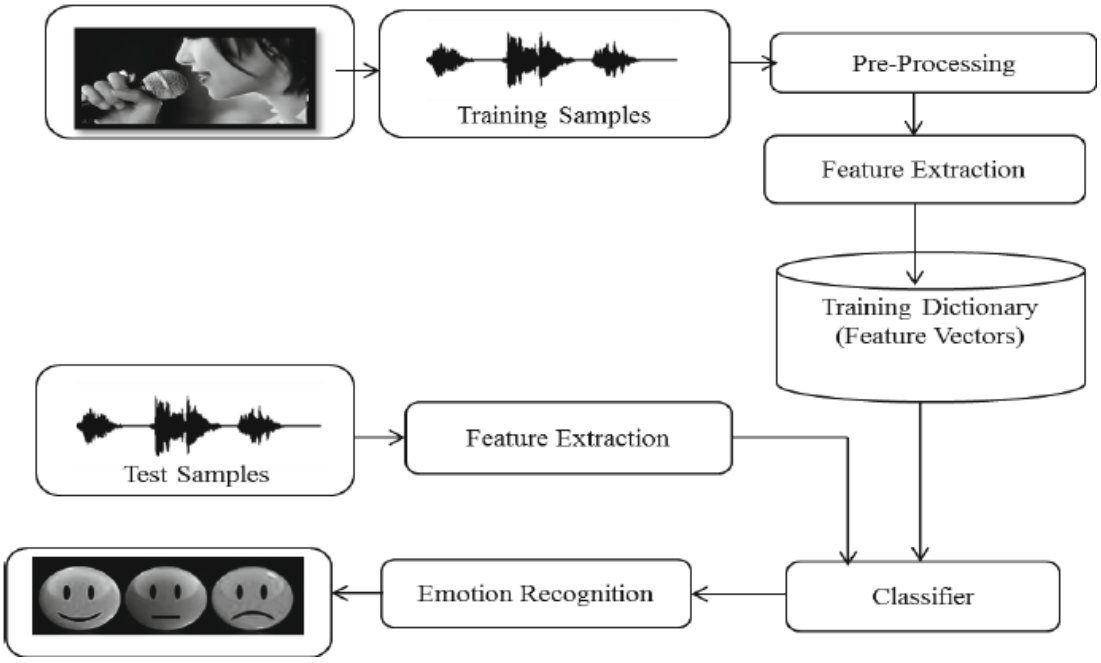

Fig. 5.1: Methodology of Telugu emotion recognition system.

5.2.1. Data acquisition

The first step in emotion recognition system is the collection of emotion database from Telugu speakers. Eight people were identified for this task and a sentence was recorded which is "**entraa ila vachhavu?**". This is a sentence in Telugu which means

"Why did you come here?" The emotions considered in this work, are **Happy, Boredom, and Neutral**. **Twenty** speech samples were recorded from each speaker using the HTC (High Tech Computer Corporation) phone for every emotion. These speeches were recorded in ".amr" format, the facility available in the smart phone. In the next step, every voice sample is converted from ".amr format" to ".mp3 format" using media software. Further the speeches are also obtained in ".wav format" using the same media software for further processing. The ".mp3 format" is very much useful for hearing the speech sentence of any speaker, whereas the ".wav format" is used in MATLAB for the proposed algorithm. There are various preprocessing steps involved to make the speech database for both training and testing samples in this proposed work. The speeches are collected from well-trained college skit artists. All the speakers are of the age group 20-25 years.

All the speeches were collected in a noiseless environment. The distance between the mike and the speaker is also maintained constant to avoid variation of the speech collected. After collecting the speech samples each speech sample was tested for quality by using Mean Opinion Score which is shown in the Table 3.1.

Each speech was heard by ten listeners, and each listener gave the rating as per their perception of the speech of particular emotion. The average of the ratings given is calculated. If MOS (Mean Opinion Score) is more than 4, then the sample is selected for the database. Otherwise the speech sample is discarded and again re-recorded.

The speech samples and their intensities of the emotions bore, Happy and neutral are shown in fig 5.2. The average time taken by the speaker, to speak in bore emotion is 2.8 seconds,

which is shown in fig 5.2(a). The time taken by the speaker to speak in Neutral emotion is 1.24 seconds as shown in fig 5.2(b) and the time taken by the speakers to speak in happy emotion is about 1.25 seconds which is shown fig 5.2(c). In general, there will be variation in the duration of speech, even for the same sentence and same speaker.

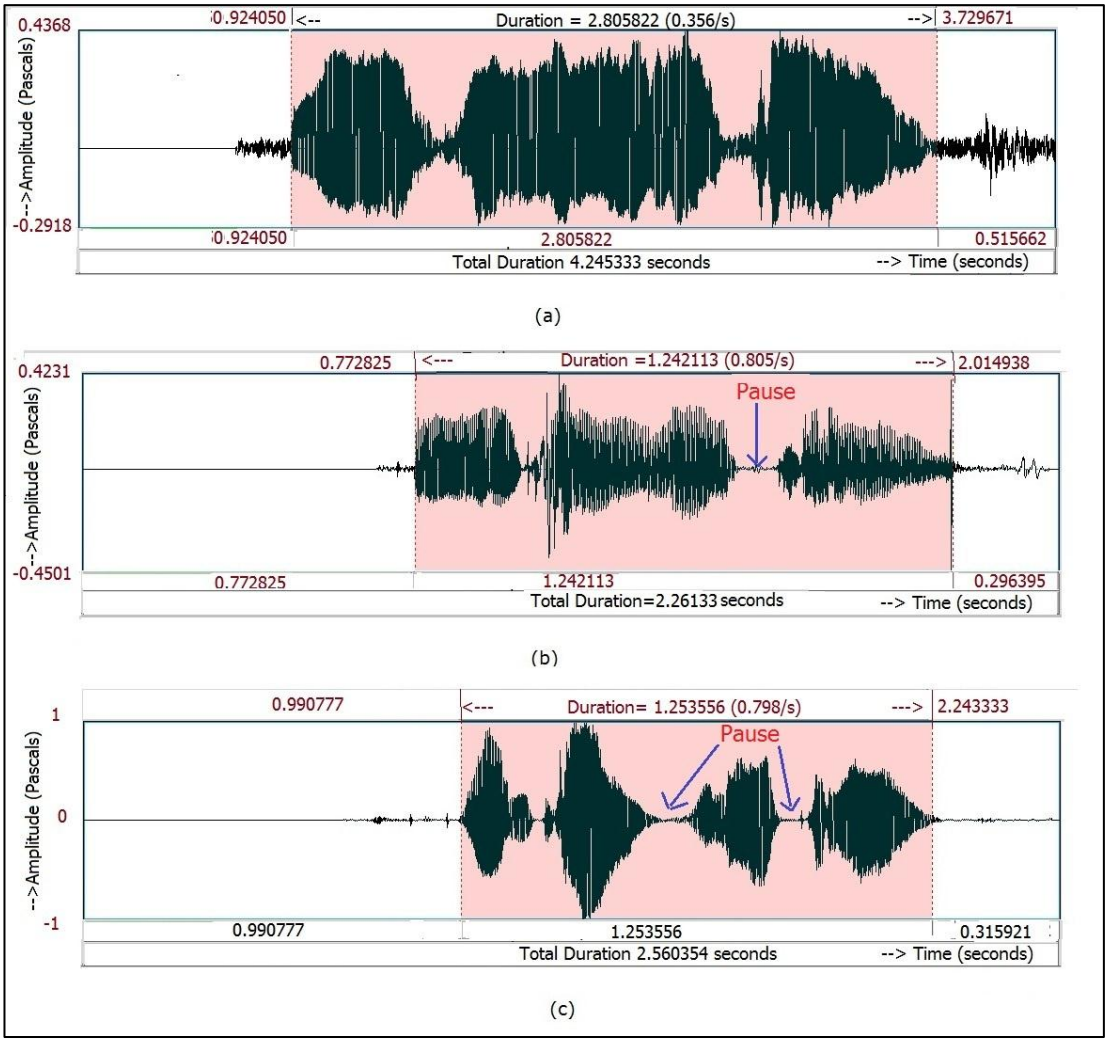

Fig 5.2: Speech signal Amplitude of emotions (a) Bore (b) Neutral (c) Happy

But when it comes to emotion speech, the pause's observed (fig 5.2(c)) in the amplitude plot represents the state of mind of the speaker. In the bore emotion, the speech duration is more due to the fact that the speakers reluctantly spoke the speech for 2.80

seconds. For the same sentence and the same speaker, the time taken in 'Happy' emotion is 1.25 seconds. Hence this feature becomes very important for classifying the "Bore" emotion from the "Happy" emotion.

The highlighted portion, which is pink in colour, is the actual speech recorded for each emotion. It is very clear from fig 5.2(c) that, there are more number of zero crossings in Happy emotion compared to Bore and Neutral. After the collection of speech databases, the relevant features like Minimum pitch, maximum pitch, average pitch, standard deviation of pitch, range of pitch, Formant F1, Bandwidth of F1, Formant F2, Bandwidth of F2, Formant F3, Bandwidth of F3, Formant F4 and Bandwidth of F4 were extracted. The test speech samples were classified using a NNC classifier as per the algorithm shown in fig 5.1. The training and testing datasets are disjoint in all our experiments.

In the next step, the testing set speeches were put into training set and the same number of training speech samples were considered as test speech samples. The algorithm given in fig 5.1 was again repeated. The recognition accuracies achieved in each fold for each emotion in each fold was obtained. The average recognition accuracy for each emotion is computed. In this way, cross validation was performed in the proposed emotion recognition work.

5.2.2. Feature extraction and Selection

After successfully creating a database, the next step in emotion recognition system is feature selection and extraction for both training and testing samples as discussed in the section 3.1.2

of chapter-3. The extracted features should have minimum distance between the samples within the same emotion class while maximize the distances between samples from the different emotion classes.

5.2.3. Classification Techniques

After the extraction of features the next step in emotion recognition system is to classify the emotions (using these extracted features) from the speech by using NNC classifier (Euclidian distance).

5.3. Results and Discussions

The Speech database was developed for various emotions like 'Happy', 'Neutral', and 'Bore'. As described in section 5.2.1, the same Telugu speech was recorded from 8 speakers. Each speaker spoke the same speech with all the 3 emotions i.e. 'Happy', 'Neutral', and 'Bore'. Each speaker spoke in 'Happy' emotion 20 times. Hence 20 samples of each emotion per person was recorded. Out of these 20 samples / emotion / speaker, 15 samples / emotion / speaker were used as a training dataset. The remaining 5 samples / emotion / speaker are used as testing dataset. Hence the total samples for the training data set becomes 360 (8 speakers X 15 Speeches X 3 emotions). Similarly for the testing dataset, the total number of speeches becomes 120 (8 speakers X 5 Speeches X 3 emotions).

The various features like Minimum pitch, maximum pitch, average pitch, standard deviation of pitch, range of pitch, Formant F1, Bandwidth of F1, Formant F2, Bandwidth of F2, Formant F3,

Bandwidth of F3, Formant F4 and Bandwidth of F4 were extracted for these speech samples. The plot for pitch of a speech for the emotions, Bore, Neutral and happy is shown in the fig 5.3. The value of the pitch for Bore emotion in fig 5.3(a) is substantially high as compared to the 'Neutral and Happy' emotion. This is due to the reason that the nasal sound **'na'** for the Telugu speech selected. The average value for the pitch of the sample for 'Bore' emotion is 284Hz.

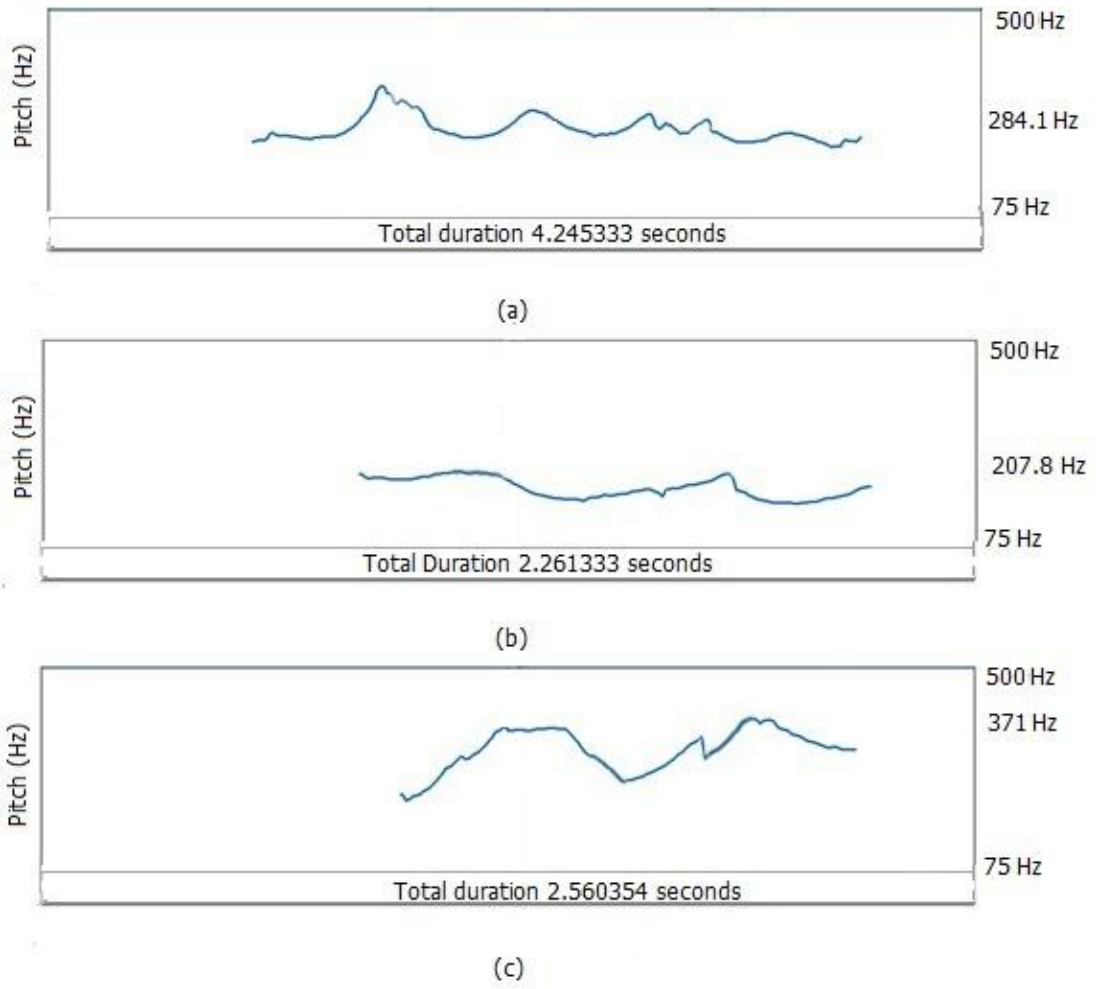

Fig 5.3: Plots of pitch vs time for emotions (a) Bore (b) Neutral (c) Happy

The value of the pitch for 'Neutral' emotion is shown in fig 5.3(b), which is almost constant for the total period of speech. The difference between the minimum and maximum pitch values for

'Neutral' emotion is very less compared to the 'Bore' and 'Happy' emotions. The average value of the pitch for 'neutral' emotion is 207Hz as shown in the fig 5.3(b).

The value of the pitch for 'Happy' emotion is shown in fig 5.3(c). The difference between the minimum and maximum values of pitch for 'Happy' emotion is very high compared to the 'Neutral' emotion. The average pitch of the speech shown in fig 5.3(c) is 371Hz.

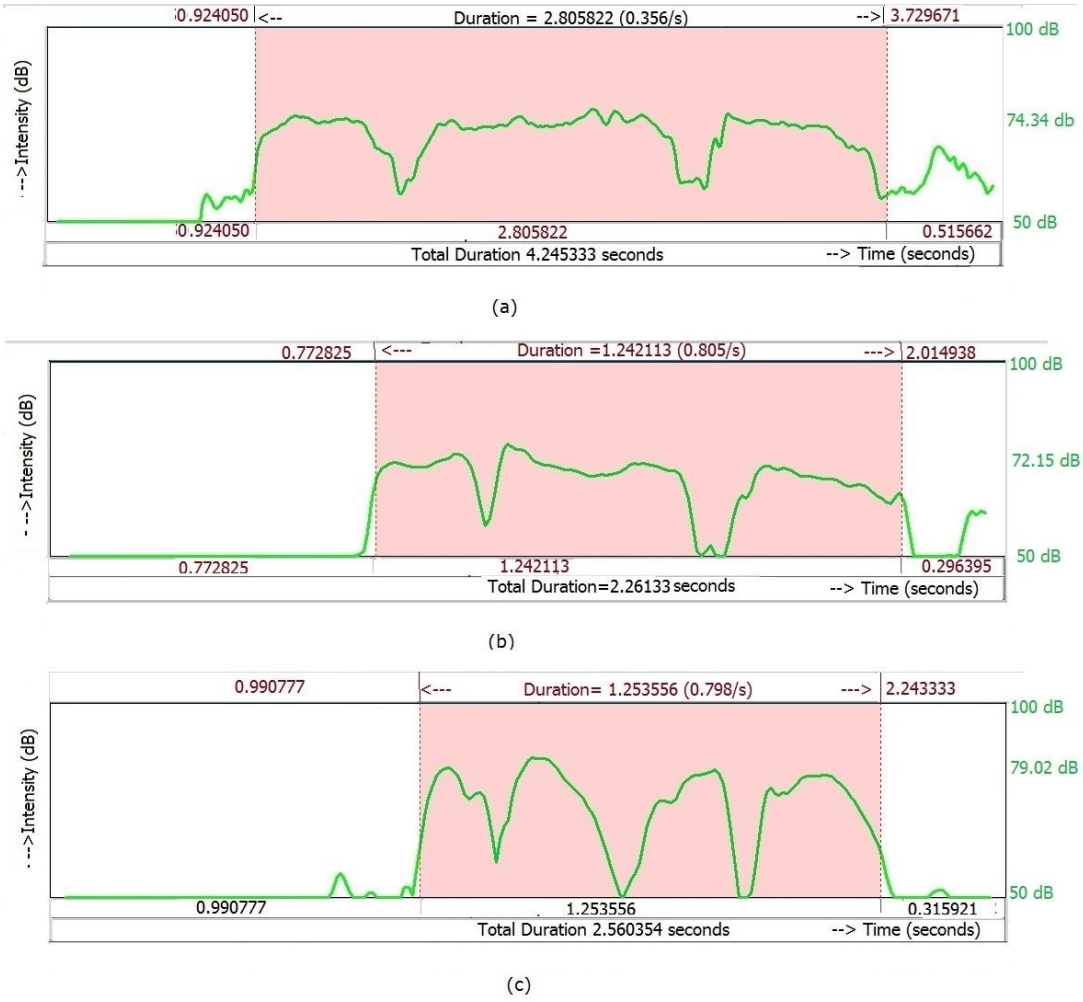

Fig 5.4: Speech signal Intensity levels of emotions (a) Bore (b) Neutral (c) Happy

The speech signal intensity levels for the emotions, 'Bore, Neutral and happy' are shown in the fig 5.4. The intensity graph for Bore emotion is shown in fig 5.4(a). The duration of this speech signal in 'Bore' emotion is found to be 2800 milli seconds as tabulated in Table 5.1. The **pauses** in the 'Bore' emotion are continued with syllables, hence the pause in the 'Bore' emotion is not clearly observed. However, the pauses in the 'Neutral and Happy' emotions are significantly observed in fig 5.4 (b) and fig 5.4 (c). When we observe and compare the intensity graphs for all these emotions, it is observed that, there are deep valleys at the time when there is a pause in sentence for 'Neutral and Happy' emotions. The average intensity value of 'Bore' emotion is 74 dB, for 'Happy' emotion it is 79 dB. This means that, the feature average intensity can be used for differentiating the 'Bore' emotion from 'Happy' emotion. The average intensity value of 'Neutral' emotion is 72 dB and that of 'Happy' emotion is 79 dB. Hence 'Happy' emotion can be easily recognised when compared with 'Neutral' emotion. The intensity plot of 'Neutral' emotion is shown in fig 5.4(b). The intensity of 'Happy' emotion is shown in fig 5.4 (c).

The average values of various features extracted from the different emotions i.e. Bore, Neutral and Happy are tabulated in Table 5.1.

The difference between the minimum and maximum value of pitch is known as the range of the pitch. The range of pitch differentiates the happy emotion very clearly from the other two emotions i.e. Bore and Neutral. The feature 'average pitch' clearly differentiates 'Happy' emotion from the 'Bore and Neutral'. The minimum, maximum, standard deviation and mean absolute slope of the pitch clearly differentiates 'Happy' emotion from the 'Bore

and Neutral'. Although, the average intensity levels for 'Bore, Happy and Neutral' are similar, the peaks observed in the intensity level for 'Happy' emotion are comparatively higher when compared with 'Bore and Neutral' emotions.

Table 5.1: Average values of Feature sets for Bore, Happy and Neutral emotions

	Bore	**Happy**	**Neutral**
Time duration of speech (milliseconds)	2800	1242	1253
Range of pitch (Hz) (Difference between minimum and maximum pitch)	169	206	132
Average pitch	224	337	216
Minimum pitch (Hz)	157	212	165
Maximum pitch (Hz)	346	422	297
Standard deviation of pitch (Hz)	32	55	25
Mean absolute slope of pitch (Hz/s)	273	526	338
Average intensity	74	79	75
Formant1 (F1) (Hz)	632	701	651
Bandwidth of F1 (B1) (Hz)	210	206	229
Formant2 (F2) (Hz)	1758	1774	1834
Bandwidth of F2 (B2) (Hz)	480	549	471
Formant3 (F3) (Hz)	2746	2775	2852
Bandwidth of F3 (B3) (Hz)	632	721	478
Formant4 (F4) (Hz)	3874	3872	3992
Bandwidth of F4 (B4) (Hz)	785	586	1962

The average value of formant F2 has a considerably high value for 'Neutral' emotion compared to 'Bore and Happy'. Hence, this feature is useful in classifying 'Neutral' emotion from 'Bore and Happy' emotions. Similarly the average value of Formant F3 is

highly useful in classifying 'Neutral' emotion from 'Bore and Happy' emotions. The average values of the Formant F4 and Bandwidth B4 are also very high for 'Neutral' emotion compared to the average values of that of 'Bore and Happy' emotions. Hence these features become very important in classifying 'Neutral' from 'Bore and Happy' emotions.

The average value of Bandwidth B3 is very high for the emotion 'Bore' compared to the average values of 'Neutral'. The duration of the speech, distinguishes the 'Bore' emotion from ' Neutral and Happy emotion.

Hence, from the table 5.1, it can be concluded that the pitch related prosodic features differentiates the Happy emotion from the other two emotions. The formant related features differentiate Neutral emotion from Happy and Bore emotion.

The system is trained with these extracted features and tested by using the NNC classifier. The result obtained for the 120 different speech samples of different emotion are tabulated in Table.5.2. The total number of samples under test for 'Neutral' emotion is 40 in number. Thirty three speech samples with 'Neutral' emotion were correctly identified in the proposed emotion recognition system. Hence the percentage of recognition accuracy for 'Neutral' emotion is [(33/40) =0.825] 82.5%. In the case of 'Happy' emotion thirty four speech samples were correctly identified, when forty samples of 'Happy' emotion were tested. Hence the percentage of recognition accuracy for 'Happy' emotion is [(34/40) =0.85] 85%. Similarly for the emotion 'Bore' twenty eight samples were correctly recognized out of forty samples tested and Hence the recognition accuracy for 'Bore' emotion

becomes[(28/40) =0.7] 70%. The overall recognition accuracy on an average for these three emotions found to be 79% ([33+34+28]/120).

Table.5.2. Result of Telugu emotion recognition system

Emotion	**No of Samples Tested**	**Correctly Identified Samples**	**Percentage of accuracy**
Neutral	40	33	82.5
Happy	40	34	85
Bore	40	28	70
Overall Recognition accuracy= 79%			

The biggest disadvantage in this area of work is non-existence of standard speech database for Indian languages [5]. Hence speech database is to be developed in laboratory for both testing and training, taking precautions to avoid background noise. This needs high amount of Zeal and motivation.

5.4. Comparison of proposed and existing approaches

The comparison between the proposed and published emotion recognition system is shown in table.5.3.

K Sreenivasa Rao and Shashidhar G Koolagudi [97] proposed Hindi speech emotion recognition system. The length of the speech is five seconds in any of the emotion for the proposed method. However the speech length in the published method is five to ten

minutes. The features selected in the proposed method are Formants, Bandwidths, Energy, intensity and Prosodic Features whereas for the published method these are Mel Frequency Cepstrum Coefficient (MFCC), Prosodic features, Energy contours. The classifier used in the published method is AANN (Auto associative Neural Network) and SVM (Support Vector machine), whereas in the proposed method Nearest Neighborhood Classifier (NNC) i.e. Euclidian distance is used.

Table.5.3: Comparison of existing and the proposed systems of emotion Recognition

Description	**Published Method**	**Proposed Method**
Language	Hindi	Telugu
Speech recorded	15 sentences in 8 given emotions	A Telugu sentence "Entra ila vachavu" in three emotions
Speech length	Five to ten minutes	Recorded for five seconds
Features Selected	Mel Frequency Cepstrum coefficients (MFCC), Prosodic features, Energy contours	Formants, Bandwidths, Energy and Prosodic Features.
Classifier	AANN (Auto associative Neural Network), SVM (Support Vector machine)	Nearest neighborhood Classifier
Recognition accuracy	78%	79%

The recognition accuracy in the proposed method is found to be 79% which is in-line with published Hindi speech emotion

recognition system. Many of the researchers are working together to develop a common speech database for different Indian languages. Further, they would be used for testing various algorithms. This would help the research to improve in future.

5.5. Conclusions

1. Due to the non-existence of emotion based Telugu speech database, a database for both training and testing was developed in the laboratory successfully.
2. The different emotions considered in this work were 'Happy', 'Neutral' and 'Bore' from 8 speakers with 20 samples for each emotion.
3. Prosodic features like minimum, maximum, range, standard deviation and mean absolute slope of pitch and intensity were extracted successfully.
4. Also other features include energy, first four formants (F1, F2, F3 and F4) and their bandwidths which were successfully extracted in this work.
5. The recognition accuracy for 'Happy' emotion is found to be 85%. The recognition accuracy for 'Neutral' emotion is found to be 82.5%. The recognition accuracy for 'Bore' emotion is found to be 70%.
6. The overall recognition accuracy of emotion recognition system by using the above features and nearest neighborhood classifier (NNC) is found to be **79%** on an average.