Emotion Recognition on The Basis of Audio Signal Using Naive Bayes Classifier

Sagar K. Bhakre
Department of E& TC
Vishwakarma Institute of Information Technology,
Pune, Maharashtra, India
E-mail:sagarbhakre@live.com

Prof.Arti Bang
Department of E& TC
Vishwakarma Institute of Information Technology,
Pune, Maharashtra, India
E-mail:-arti.bang@viit.ac.in

***Abstract*— In this paper we have studied and implemented the classification of audio signal into four basic emotional state. For that we have considered different statistical features of pitch, energy, and ZCR (Zero Crossing Rate) MFCC (Mel frequency cepstral coefficient) from 2000 utterances of the created audio signal database. In that, Pitch feature is extracted by AMDF (average magnitude difference method) and energy is calculated by sum of square absolute value of magnitude spectrum. And MFCC is calculated by taking DCT (Discrete cosine transform) of its energies spectrum by keeping the DCT coefficients 1-14 and discarding the rest. In statistical modeling, regression analysis is a statistical process for calculating approximately the variables. It comprise many techniques for modeling and analyzing several variables. In this paper Naïve Bayes Classifier is used to classify the audio signal into four different emotions. Speech signal is random signal so we have to predict the future sample and Naïve Bayes Classifier is totally probability based classifier so in speech analysis for accurate prediction we are using Naïve Bayes classifier. In the speech signal for recognition of signal classifier require millions of dataset. The advantage of Naïve Bayes classifier is that it recognizes the signal with minimum dataset**

Keywords*— *pitch, energy, MFCC, ZCR, Naïve Bayes Classifier.

I. Introduction

Emotion recognition is a technique used to identify the different emotions of human beings using their speech signal. Research community is now attracting towards the new area of interest that is speech emotion recognition. To normalize the problem of automatic speech recognition, Speech emotion classification can be used. Furthermore, emotion and influence information were extracted from speech which is useful to enhance human-computer interaction. The proposed systems is classifying the speech signal into four classes i.e. anger, happy, sad, neutral. The global statistical features of energy, pitch, and MFCC (Mel frequency cepstral coefficient), ZCR (Zero crossing rate), formant, LPC (Linear predictive coding) have been important for emotion classification [1] [2] [4]. The proposed system evaluates and classifies statistical features of energy, pitch, MFCC, ZCR.

The system is implemented with respect to the probability of accurate classification which have been achieved by the Naive Bayes classifier that models the pdf of features as a mixture. The criterion employed for electing the best features is the probability of correct classification that is determined via cross validation [3]. The initiate approach for the classification of speech is different for male and female.

II. Methodology

The proposed methodology for emotion recognition system is shown in fig.1 [9]

A. Dataset characterization

We have created a dataset of audio signal for analysis. Dataset consist of 2000 sentences from 20 different speakers in the age group of 18 to 30 years. The mother tongue of these speakers is Marathi, Each of this speakers uttered 100 sample of all the emotion. The dataset language was chosen to be English and these sentences were spoken and recorded as naturally as possible in anger, happy, sad, and neutral respectively. The audio signals are recorded in control environment area at sampling frequency 16K Hz by Plantronics headphone. The average length of each sentence is 2 to 4 seconds depending on the emotion. The dataset is about 65 to 70 min in duration

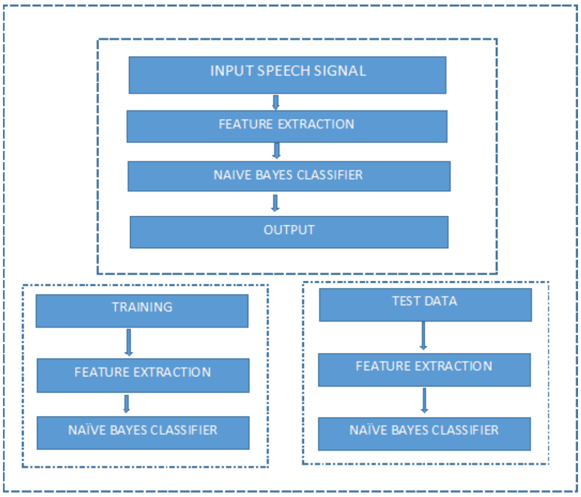

Fig.1. Emotion recognition system

B. Feature extraction

To evaluate the emotions from the audio signal one has to carefully select suitable features. First, we have to carry information about the transmitted audio signal. Therefore we need to choose classification algorithms for further classification. Hence we aim to choose created dataset statistic and compare with the instantaneous signal statistic, for that we have to select appropriate feature. For good comparability feature we need to derive by the same underlying contours. Firstly we have to derive. .the raw contour and on the bases of this we have to derive the adapted feature vectors

We choose the analysis of the contours of pitch, energy, MFCC and we are well known about their capability by considering a large amount of user's emotion information. The selected contours are broadly based on sound while spectral characteristic depends on their phonemes and therefore we have to consider the phonetic content of an utterance. This is a main drawback of thinking of independency of the spoken content or even the language. In order to calculate the contours, frames of the speech signal are analyzed every 15ms using a Hamming window function. By using hamming window frame of speech signal are analyzed every 15ms in order to calculate the counters by observing a signal for every 15ms using hamming window we have calculated a contour and frames of audio signal as shown in figure 2.

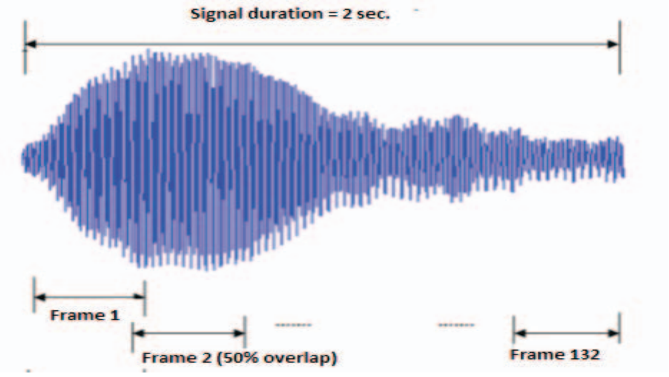

Fig.2 Recorded original audio signal.

1) Spectral features

a) *Pitch features*

In general, pitch features has more impact than energy feature for classification. The near periodic vibration of vocal folds is excitation for the production of voice speech. Extracted different statistical features of pitch are shown below [1] [2].

- Mean of pitch.
- Median of pitch.
- Standard deviation of pitch.
- Corresponding pitch maxima.
- Corresponding pitch minima.

b) Intensity (Energy) features

Alike pitch features, Energy feature can also contribute to distinguish between the emotions of audio signal. Extracted different statistical features of energy are shown below [3].

- Mean of energy.
- Median of energy.
- Standard deviation of energy.
- Corresponding energy maxima.
- Corresponding energy minima

c) Zero-crossing rate (ZCR)

It is the rate at which the signal changes from positive to negative or back. This feature has been used heavily in both speech recognition and emotion recognition being an important feature to classify the audio signal.

- Mean of ZCR.
- Median of ZCR.
- Standard deviation of ZCR.
- Corresponding ZCR maxima.
- Corresponding ZCR minima

d) Mel-Frequency Cepstral Coefficients (MFCC)

The block diagram for calculation of different MFCC coefficients is shown below in figure 3. [8]

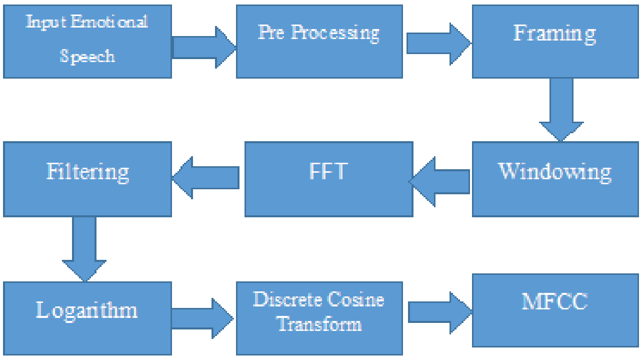

Fig.3 Block diagram for calculation of MFCC

In the proposed system MFCC feature are extracted from input speech signal [4] [5]. MFCC is used to create fingerprint of input speech signal. To detect the important characteristics of speech MFCC use human's ear's critical bandwidth frequencies variation with filters spaced at logarithmically at high frequencies and linearly at low frequencies. Human perception about frequency of audio signal does not follow the linear scale so in that mel scale is used to measure the pitch. In the mel scale below 1000 Hz. there is linear spacing frequency above that logarithmic spacing. The formula is used to compute the mels for a frequency mel (f) = 2595*log10 (1+ f / 700) [4]. A. To remove the acoustical interference in the speech waveform is cropped. Tapping a zero at the beginning and end of every frame window block reduces the discontinuities of speech signal. The FFT block convert each frame into frequency domain from time domain.

Mel filtering block, speech signal is plotted mel spectrum to mimic hearing. Finally standard frequency scale is created form mel spectrum scale. This spectral properties is key for recognizing characteristic of the signal. Once the fingerprint is created which is referred as an acoustic vector. These vectors retains 12-15 DCT coefficient. The 0th coefficient is discarded because it corresponds the energy of the frame. The resultant are Mel-Frequency Coefficients out of which 14 MFCC coefficient are calculated, which are the remaining extracted 14 features.

C. Classification using Naïve Bays Classifier

The Naïve bayes is a conditional probability model in that a problem instance vector $x = (x_1, x_2, \dots \dots x_n)$ in our case x is the values of extracted features. The probability model of Navie bayes classifier is shown in equation 1.

$$p(A_k|x) = \frac{p(A_k)\, p(x|A_k)}{p(x)} \qquad (1)$$

Where k = possible outcome or classes A_k

Thus the joint probability model can be expressed as

$$p(A_k|x_1, x_2, .. x_n) \propto p(A_k) \prod_{i=1}^{n} p(x_i|A_k) \qquad (2)$$

Automatic recognition is the machine learning technique which require trained dataset which is having a collection of emotional speech signal recorded in different emotion [7] [8]. Each recorded signal in the dataset has labelled with the different emotional state. A classifier is made by trained dataset and algorithm is form on the basis of the dataset. Once a train dataset is perfectly classified under four different headings, such as anger, happy, sad, and neutral. Naive based classifier can able to predict the emotion which is not in the dataset once it is train. On the basis of feature i.e. pitch, energy, MFCC we created a dataset and on the basis of that we are classifying the speech signal

In Naïve Bayes classifier we used its default setting of classifier. Speech corpus [8]. The classification systems were first implemented using pitch, MFCC and Energy feature. The confusion matrix of training data for classification system in terms of the average percentage of recognition of the 4 different emotional states in the database using the various aggregates [6]. All the test were done on same dataset.80% of data is used as a dataset while reaming data used as test sample to check the result

III. EXPERIMENTATION AND RESULTS

Created database is of recorded utterance of male between age of 18-30 is about length of 2sec each. We have taken a total of 2000 utterances of four emotions state and each consist of 500 samples. Once the sample dataset is created it undergoes preprocessing. Speech signal comprises with silence, Unvoiced and voice part. For good recognition it is important to remove the silence part. To remove it minimum mean squared method is used. In that threshold is fix at 0.03, fs=16000 Hz, index=1000 as shown in fig 4(a) and 4(b).Then the signal undergoes feature extraction. In the given paper we have extracted pitch by using AMDF method while energy feature is extracted taking square magnitude spectrum and MFCC is calculated by taking DCT of its energies spectrum by keeping the DCT coefficients 1-14 and discarding the rest. Then this feature are consider as their statistical value like mean, standard deviation, median, Minimum and Maximum values. The distribution of pitch's mean, energy's mean, ZCR's mean and single MFCC coefficient over the dataset is sown in figure 5, 6, 7 and 8 respectively. By considering the above extracted features we have created training dataset consist of 90% of total sample i.e. 1800 sample and 10% for testing. For classification of test dataset we have used Naïve Bayse classifier [8] whose result is shown as in the form of confusion matrix in table I and number of correctly classified sample is shown in table II. We get baseline result with 10-fold cross validation. The above mentioned experimentation is successfully done in MATLAB 2015a platform.

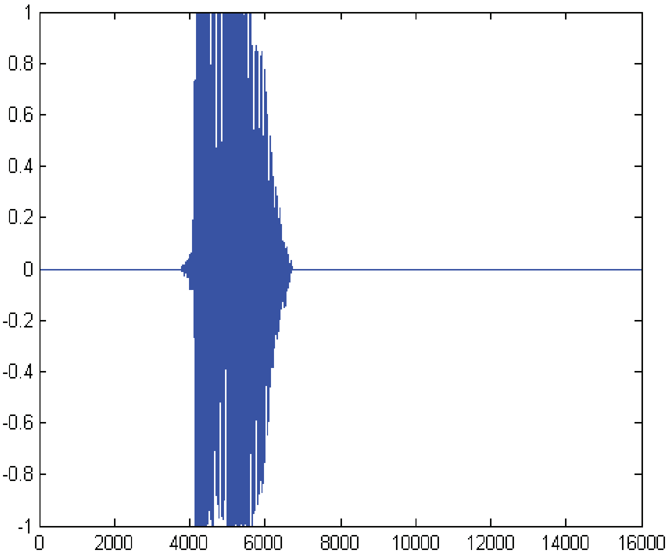

Fig. 4(a) Original recorded signal

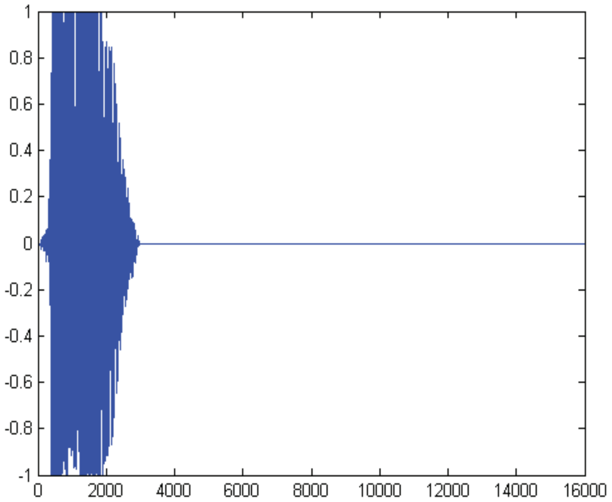

Fig. 4(b) preprocess recorded signal where silence and unvoiced part is tends to zero

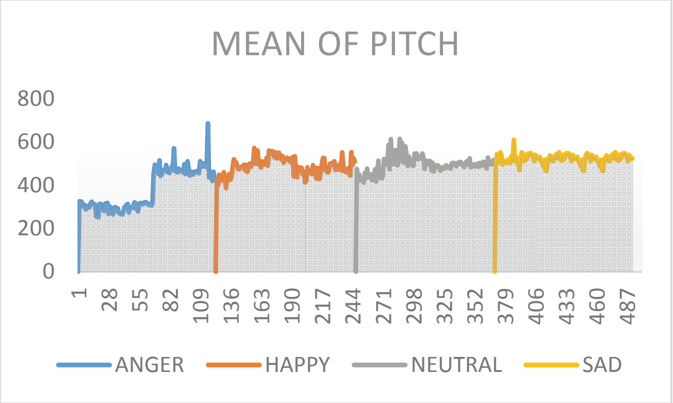

Fig. 5 Mean of Pitch distrubution over the created dataset

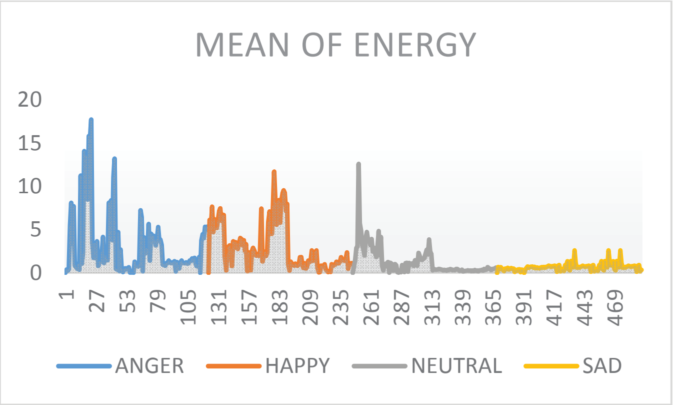

Fig. 6 Mean of Energy distrubution over the created dataset

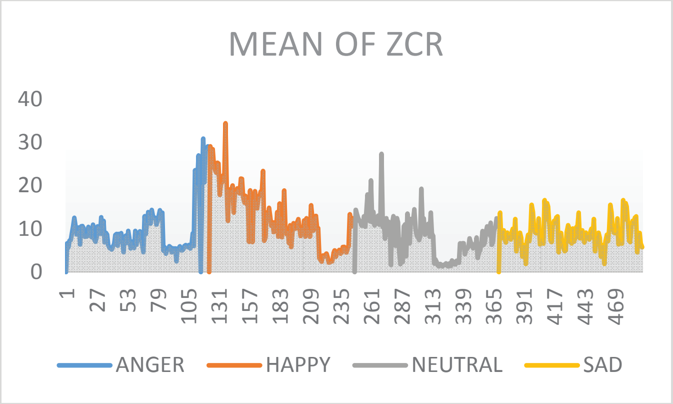

Fig. 7 Mean of ZCR distrubution over the created dataset

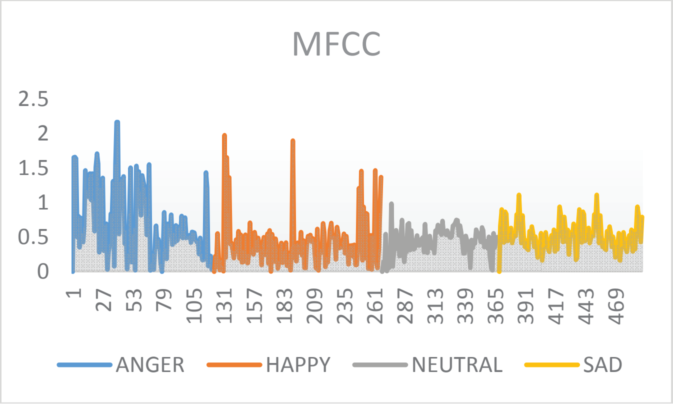

Fig. 8 Single MFCC coeficient distrubution over the created dataset

- accuracy for happy is 78%
- accuracy for anger is 81%
- accuracy for neutral is 77%
- accuracy for sad is 76%

Table.1. Confusion Matrix of a Naïve Bayes classifier

	RESPONSQE			
EMOTION	ANGER	HAPPY	SAD	NEUTRAL
ANGER	39	10	0	1
HAPPY	8	40	1	1
SAD	0	1	39	10
NEUTRAL	0	0	12	38

Table.2. Classification Results.

SR	EMOTIONAL INPUT	RESPONSE	
NO.	SENTENCE	NO. OF INPUT SAMPLE	CORRECT OUTPUT RESULT
1	ANGER	50	39
2	HAPPY	50	40
3	SAD	50	39
4	NEUTRAL	50	38

IV. CONCLUSION

In this paper, a complete speech-based emotion recognition framework is done using naïve Bayes classifier. Naïve Bayes classifier is trained by the extracted feature such as pitch, energy, MFCC. In this system we achieved accuracy for anger happy, sad, neutral 81%, 78%, 76%, 77% respectively. The results extracted and it show that the proposed system is capable of real-time emotion recognition. Working with the set of utterances, the highest probability of correct classification for is achieved for Naive Bayes Classifier. For male utterances, separate speech classifiers were built. Feature extraction methods are the main area of study in the proposed system which are useful in emotion recognition. Most recognizable features include the pitch, the short term energy and the MFCCs. For future work, to enhance the system performance proposed system can be configured to work with voice quality and prosodic features

REFERENCES

[1] A. A. Khulage, "Extraction of pitch, duration and formant frequencies for emotion recognition system," *Communication and Computing (ARTCom2012), Fourth International Conference on Advances in Recent Technologies in*, Bangalore, India, 2012)

[2] A. Agarwal and A. Dev, "Emotion recognition and conversion based on segmentation of speech in Hindi language," *Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on*, New Delhi, 2015, pp. 1843-1847..

[3] S. H. Chen, Y. S. Lee, W. C. Hsieh and J. C. Wang, "Music emotion recognition using deep Gaussian process," *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Hong Kong, 2015R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[4] M. T. S. Al-Kaltakchi, W. L. Woo, S. S. Dlay and J. A. Chambers, "Study of fusion strategies and exploiting the combination of MFCC and PNCC features for robust biometric speaker identification," *2016 4th* International Conference on Biometrics and Forensics (IWBF)*, Limassol, 2016, pp. 1-6.*

[5] K. V. Krishna Kishore and P. Krishna Satish, "Emotion recognition in speech using MFCC and wavelet features," *Advance Computing Conference (IACC), 2013 IEEE 3rd International*, Ghaziabad, 2013, pp. 842-847.

[6] M. Durukal and A. K. Hocaoğlu, "Performance optimization on emotion recognition from speech," *2015 23nd Signal Processing and Communications Applications Conference (SIU)*, Malatya, 2015, pp. 308-311.

[7] Ç Oflazoglu and S. Yıldırım, "Anger recognition in Turkish speech using acoustic information," *2012 20th Signal Processing and Communications Applications Conference (SIU)*, Mugla, 2012, pp. 1-4.

[8] N. Sebe, M. S. Lew, I. Cohen, A. Garg and T. S. Huang, "Emotion recognition using a Cauchy Naive Bayes classifier," *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, 2002, pp. 17-20 vol.1.

[9] Vinay a, Shilpi Gupta b, Anu Mehra c, " Vocal Emotion Recognition using Naïve Bayes Classifier," *Proc. of Int. Conf. on Advances in Computer Science, AETACS*, 2013.